UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANÁLISE EXPLORATÓRIA E COMPARATIVA DA APLICAÇÃO DE AGRUPAMENTO PARA COMBATE À LAVAGEM DE DINHEIRO

Dissertação de Mestrado

FABIO MANGUEIRA DA CRUZ NUNES

SÃO CRISTÓVÃO/SE

2019

UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FABIO MANGUEIRA DA CRUZ NUNES

ANÁLISE EXPLORATÓRIA E COMPARATIVA DA APLICAÇÃO DE AGRUPAMENTO PARA COMBATE À LAVAGEM DE DINHEIRO

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação (PROCC) da Universidade Federal do Sergipe (UFS) como parte de requisito para obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Methanias Colaço Rodrigues Júnior

SÃO CRISTÓVÃO/SE
2019

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

N972a

Nunes, Fábio Mangueira da Cruz
Análise exploratória e comparativa da aplicação de agrupamento para combate à lavagem de dinheiro / Fabio Mangueira da Cruz Nunes ; orientador Methanias Colaço Rodrigues Júnior. - São Cristóvão, 2019.
94 f. : il.

Dissertação (mestrado em Ciência da Computação) – Universidade Federal de Sergipe, 2019.

1. Computação. 2. Lavagem de dinheiro. 3. Mineração de dados (Computação). 4. Algoritmos computacionais. I. Rodrigues Júnior, Methanias Colaço orient. II. Título.

CDU 004.056:343.3

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Ata da Sessão Solene de Defesa da Dissertação do Curso de Mestrado em Ciência da Computação-UFS. Candidato: FÁBIO MANGUEIRA DA CRUZ NUNES

Em 31 dias do mês de janeiro do ano de dois mil e nove, com início às 15h00min, realizou-se no miniauditório do CCET da Universidade Federal de Sergipe, na Cidade Universitária Prof. José Aloisio de Campos, a Sessão Pública de Defesa de Dissertação de Mestrado do candidato **FÁBIO MANGUEIRA DA CRUZ NUNES**, que desenvolveu o trabalho intitulado: ***"Análise Exploratória e Comparativa da Aplicação de Agrupamento para Combate à Lavagem de Dinheiro"***, sob a orientação do Prof. Dr. METHANIAS COLACO RODRIGUES JUNIOR. A Sessão foi presidida pelo Prof. Dr. METHANIAS COLACO RODRIGUES JUNIOR (PROCC/UFS), que após a apresentação da dissertação passou a palavra aos outros membros da Banca Examinadora, Prof. Dr. ADICINÉIA APARECIDA DE OLIVEIRA (PROCC/UFS) e, em seguida, ao Prof. PAULO CAETANO DA SILVA(UNIFACS). Após as discussões, a Banca Examinadora reuniu-se e considerou o mestrando (a) APROVADO "(aprovado/reprovado)". Atendidas as exigências da Instrução Normativa 01/2017/PROCC, do Regimento Interno do PROCC (Resolução 67/2014/CONEPE), e da Resolução nº 25/2014/CONEPE que regulamentam a Apresentação e Defesa de Dissertação, e nada mais havendo a tratar, a Banca Examinadora elaborou esta Ata que será assinada pelos seus membros e pelo mestrando.

Cidade Universitária *"Prof. José Aloisio de Campos"*, 31 de Janeiro de 2019.

Prof. Dr. Methanias Colaço Rodrigues Júnior
(PROCC/UFS)
Presidente

Prof. Dr. Adicinéia Aparecida de Oliveira
(PROCC/UFS)
Examinador Interno

Prof. Dr. Paulo Caetano da Silva
(UNIFACS)
Examinador Externo

Fábio Mangueira da Cruz Nunes
Candidato

A meu falecido Pai, o meu maior incentivador, meu melhor amigo.

Agradecimentos

Aos meus pais, por terem me mostrado o valor da educação, justiça, verdade e fé, pelo amor incondicional e por estarem sempre ao meu lado.

Ao meu amor Deyseany Nunes pelo apoio, compreensão e paciência. Sem você ao meu lado não conseguiria concluir mais esse desafio.

As minhas filhas lindas que amo, Mariana Lima e Anna Luisa Lima, por quem continuo lutando.

Ao meu orientador, Prof. Dr. Methanias Colaço Rodrigues Júnior, por acreditar, meu eterno mestre, um amigo, um parceiro, verdadeiramente um professor, a quem verdadeiramente tenho como referência acadêmica! Um irmão onde encontrei abrigo.

Aos amigos e companheiros de mestrado: Rafael e Othon, e todos os demais que contribuíram diretamente.

Ao amigo Bruno Kreuts, você é um parceiro fantástico, obrigado pela ajuda e apoio.

Obrigado a todos que, de alguma forma, contribuíram para o meu crescimento pessoal e para a realização deste trabalho.

RESUMO

Contexto: Desde 2007, por meio da Estratégia Nacional de Combate à Corrupção e à Lavagem de Dinheiro (ENCCLA), iniciou-se a criação dos primeiros Laboratórios de Tecnologia contra Lavagem de Dinheiro (LABLDs), os quais, hoje, estão presentes em todas as regiões da federação e são responsáveis por políticas de desenvolvimento de métodos e tecnologias de ponta para dar suporte aos órgãos de persecução penal. A necessidade de inovação neste cenário de combate ao crime impõe parcerias, apoio, pesquisas e método científicos. **Objetivo:** Este trabalho teve por proposito avaliar a eficácia dos algoritmos EM (*Expectation–Maximization*) e *K-Means* sobre bases de dados reais de transações financeiras investigadas pelos LABLDs de Sergipe, comparando as evidências encontradas com os resultados obtidos pelo mapeamento do estado da arte publicado na literatura. **Método:** Inicialmente, foi realizado um Survey com a premissa de caracterizar a utilização de técnicas de armazenamento, integração, Data Mining e Data Analytics pelos LABLDs e demais unidades investigativas em todo o Brasil. Em seguida, foi executado um mapeamento sistemático como forma de identificar e sistematizar as principais abordagens, técnicas e algoritmos usados na computação, para lutar contra a LD. Por fim, foi planejado e executado um experimento controlado, in vivo, para comparar os algoritmos EM e K-Means. **Resultados:** Constatou-se que aproximadamente 97% dos respondentes do survey não utilizavam diretamente algum algoritmo de mineração de dados e que 30,99% avaliavam o próprio conhecimento sobre o assunto como ruim ou péssimo. Para o estado da arte, foi identificado que as abordagens principais utilizadas contra LD são classificadores supervisionados e clusters. Com a execução do processo experimental, foi evidenciado que o algoritmo EM supera o algoritmo K-means, alcançando uma acurácia média máxima de 98,25%. **Conclusões:** Esta dissertação expôs a realidade dentro dos principais órgãos de investigação e controle do nosso país. Após ser analisado o estado da arte, evidenciou-se que há oportunidades para explorar soluções contra LD, principalmente nas áreas de Aprendizado de Máquina e Aprendizado Profundo. Finalmente, o algoritmo EM se apresentou como uma alternativa superior ao K-means, para a implementação de um módulo preditor de transações suspeitas, confirmando os resultados da literatura, todavia, em um ambiente real e específico de investigação.

Palavras-chave: Lavagem de Dinheiro; Data Mining; Data Analytics; K-Means; Expectation–Maximization (EM).

ABSTRACT

Context: Since 2007, through the National Anti-Corruption and Money Laundering Strategy (ENCCLA), the first LABLDs have been created, which are present now in all the regions of the federation and are responsible for policies to develop methods and advanced technologies to support the bodies of criminal prosecution. The need for innovation in this crime combat scene imposes partnerships, support, research and scientific method. The **objective** of this work was to evaluate the effectiveness of the Expectation-Maximization (EM) and K-Means algorithms on real financial transaction databases investigated by Sergipe's LABLDs, comparing the evidences found with the results obtained by mapping the state-of-the-art published in the literature. **Method**: Initially, a Survey was conducted with the premise of characterizing the use of techniques of storage, integration, Data Mining and Data Analytics by LABLDs and other investigative units throughout Brazil. Then, a systematic mapping was performed as a way to identify and systematize the main approaches, techniques and algorithms used in computer science to combat LD. Finally, a controlled in vivo experiment was designed and executed to compare the EM and K-Means algorithms. **Results**: It was found that approximately 97% of survey respondents did not directly use any data mining algorithm and that 30.99% evaluated their own knowledge about the subject as bad or very bad. Related to the state of the art, it has been identified that the main approaches used against LD are supervised classifiers and clusters. With the execution of the experimental process, it was evidenced that the algorithm EM surpasses the K-means algorithm, reaching a maximum average accuracy of 98.25%. **Conclusions**: This thesis exposed a hard reality within the main investigation and control bodies of our country. After analyzing the state of the art, it was evidenced that there are opportunities to explore solutions against LD, especially in the areas of Machine Learning and Deep Learning. Finally, the EM algorithm presented as a superior alternative to K-means for the implementation of a predictor module of suspicious transactions, confirming the results of the literature, however, in a real and specific investigation environment.

Keywords: Money Laundering; Data Mining; Data Analytics; K-Means; Expectation-Maximization (EM).

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

BI	*Business Intelligence*
DRCI	*Departamento de Recuperação de Ativos e Cooperação Jurídica Internacional*
DW	*Data Warehouse*
ETL	*Extraction, Transformation and Load*
EM	*Expectation Maximization*
EMV	*Estimador de Maxima Verissimilhança*
GAECO	*Grupo de Atuação Especial de Combate ao Crime Organizado*
LABLD	*Laboratório de Tecnologia Contra a Lavagem de Dinheiro*
LD	*Lavagem de Dinheiro*
OLAP	*On-line Analytical Processing*
ISP	*Inteligência de Segurança Pública*
RMSE	*Root Mean Square Error*

SUMÁRIO

CAPÍTULO 1

1. INTRODUÇÃO

1.1 Contextualização

No Brasil, os órgãos especializados em investigações complexas têm o papel fundamental de produzir conhecimento e informações de valor significativo, podendo assim subsidiar as investigações com grande volume de dados e trazer luz aos fatos contidos nos afastamentos de sigilos quebrados por decisão judicial, além dos que estão disponibilizados na web e que atendam às necessidades investigativas dos processos (Brasil, 1999). Contudo, a atividade de Inteligência de Segurança Pública (ISP) merece um destaque especial, pois deverá exercer sempre sua função básica de produzir conhecimento de interesse e utilidade para a instituição policial ou para outro órgão de controle que também exerça atividade correlata à investigação policial.

Desta forma a Lei 9883/1999, de 07 de dezembro de 1999 onde surge o Sistema Brasileiro de Inteligência define :

> "a atividade que objetiva a obtenção, análise e disseminação de conhecimentos dentro e fora do território nacional, sobre fatos e situações de imediata ou potencial influência sobre o processo decisório e a ação governamental, e sobre a salvaguarda e a segurança da sociedade e do Estado"

Nesta corrente, a Doutrina Nacional de Inteligência de Segurança Pública também aporta, definindo a ISP como:

> "o exercício permanente e sistemático de ações especializadas para identificar, avaliar e acompanhar ameaças reais ou potenciais na esfera de Segurança Pública, basicamente orientadas para produção e salvaguarda de conhecimentos necessários para subsidiar os tomadores de decisão, para o planejamento e execução de uma política de Segurança Pública e das ações para prever, prevenir, neutralizar e

> reprimir atos criminosos de qualquer natureza que atentem à ordem pública, à incolumidade das pessoas e do patrimônio".(Brasil, 2014)

Baseados nesta Doutrina, quando nos deparamos com um universo de ações criminosas que são executadas diariamente em nossa sociedade, para as quais o nosso mecanismo de defesa e proteção é a segurança pública e demais órgãos de controle (Costa & Lima, 2014), percebemos que se faz relevante entender o seu meio de funcionamento, sua estrutura e os seus resultados, bem como se seus esforços têm usado técnicas automatizadas para analisar diferentes tipos de crimes, mesmo sem um arcabouço unificador que descreva como aplicá-las (Chen et al., 2014).

Dentro do campo da atividade investigativa, a investigação, na prática, nada mais é do que uma busca constante por um grande volume de dados, de maneira a levar o investigador, no sentido amplo da palavra, à identificação e ao evidenciamento de informações que possam trazer à luz dados concretos sobre o fato investigado (Santos, 2012). Mais especificamente, é a busca por informações que possam levar os investigados aos rigores da lei.

Essa realidade prática impõe uma dependência das informações e dos dados disponíveis em diversas bases, sejam estas de fontes primárias ou secundárias, perfazendo um fluxo de informações que necessita de um trabalho diferenciado, a ser exercido por profissionais capazes de produzir os resultados esperados.

Diante dessa problemática existente nas atividades policiais, um modelo de bases de dados ideal deve abranger dados dos órgãos estaduais e federais (Figueira, 2015), para que haja efetividade na busca por resultados e soluções. Este modelo exigirá um conjunto de ações que possibilitem a sistematização do fluxo de dados e desburocratização em torno da comunicação entre as agências, de modo que as informações trafeguem numa constância ininterrupta e contribuam para a solução de crimes.

Desta forma, fica notória a necessidade de integração de dados para realização de um trabalho eficiente, sem ser desconsiderada a complexidade desta

atividade, uma vez que os dados precisam ser modelados e os metadados padronizados, além da exigência de um ambiente específico para processamento e análise, o qual está diretamente relacionado com a qualidade dos dados que são armazenados numa base histórica (Bramer, 2007; Kimball et al, 2011).

Essa base histórica tende a ser volumosa e complexa, com características do fenômeno Big Data (Witten et al., 2016), as quais posicionam o processo investigativo como desafiador e ávido por técnicas que auxiliem as suas atividades. Neste sentido, para assessorar na extração de conhecimento, técnicas de Data Mining (Mineração de Dados) e Data Analytics (Análise de Dados) são abordagens muito utilizadas para descobrir padrões e extrair informações que podem ser úteis a tarefas de investigação (Mcafee et al., 2012).

Assim, percebemos quão complexa é a análise criminal e os desafios do analista criminal, principalmente quanto às análises que envolvem os crimes mais complexos[1], a exemplo da Lavagem de Dinheiro.

Neste contexto, a apreciação dos insumos originários das transações financeiras para detecção dos crimes de lavagem de dinheiro, apoiadas por modelagem computacional e associadas às ferramentas corretas, algorítimos de *Data Mining* e *Data Analytics*, proporcionará aos analistas criminais um portfólio minimamente suficiente de possibilidades recursivas e investigativas, aumentando sobremaneira o poder de detecção de cada agente investigador dentro de cada cenário.

Todavia, em razão da dificuldade de encontrar uma abordagem com embasamento científico que ratifique as dificuldades dos analistas criminais em seu trabalho diário dentro dos LABLDs, bem como do contexto desafiador da análise criminal, iniciou-se a linha de pesquisa desta dissertação de mestrado.

[1] São crimes cujo volume de informações e dificuldades em acessar os dados, dificultam as investigações.

1.2 Análise do problema

Em 2007, ficou definido como meta a ser executada pela Estratégia Nacional de Combate à Corrupção e à Lavagem de Dinheiro (ENCCLA, 2006), a criação do primeiro Laboratório de Tecnologia contra Lavagem de Dinheiro (LAB-LD), por meio de cooperação técnica firmada entre o Ministério da Justiça e o Banco do Brasil, no âmbito do Departamento de Recuperação de Ativos e Cooperação Jurídica Internacional (DRCI).

Esta unidade teve como principal característica o uso e o desenvolvimento de tecnologia de ponta que pudesse servir de aporte às Polícias Judiciárias e aos Ministérios Públicos, com replicação do modelo para outros órgãos estaduais e federais.

Desta forma, vários laboratórios de tecnologia contra lavagem de dinheiro foram criados, vislumbrando contar com uma abordagem que utilizasse uma metodologia própria para análise de dados, focada nos crimes mais complexos, relacionados à corrupção, à lavagem de dinheiro e às organizações criminosas.

Neste sentido e fora do contexto brasileiro, em Chen et al. (2014), após caracterizarem o K-Means como técnica mais pesquisada para a detecção de transações suspeitas, compararam-no com o algoritmo de Maximização da Expectativa e concluíram que este supera o K-Means.

Em outro trabalho internacional, Duhart et al. (2016) descreveram brevemente que, na literatura, as detecções automáticas de atividades suspeitas compreendem o uso de técnicas classificadas como estatísticas e de inteligência artificial.

No Brasil, tratando-se da rede de Laboratórios de Tecnologia Contra a Lavagem de Dinheiro - LAB-LDs, o grande mote não é a fraude em si na sua concepção jurídica, como descrita no Art. 171, caput do código penal brasileiro, mas, na detecção de artifícios utilizados pelos criminosos, no intuito de fazer adentrar no mercado bancário dinheiro advindo do cometimento de ilícitos penais.

A questão é: como os LABLDs brasileiros utilizam abordagens computacionais para detecções de crimes de LD?

Este contexto de necessidade de entender melhor o status quo da rede LABLDs instigou a realização de um survey aplicado aos principais órgãos brasileiros de combate ao crime organizado (Santos & Mangueira et al., 2017), tais como as agências de Inteligência de Segurança Pública – ISP, os Laboratórios de Tecnologia Contra a Lavagem de Dinheiro – LAB-LDs - e os Grupos de Atuação Especial de Repressão ao Crime Organizado – GAECO, com o objetivo de conhecer o cenário atual da utilização de ferramentas de Data Analytics e de Data Mining nestas agências (https://pt.surveymonkey.com/r/PesquisaBI). Este survey respondeu às seguintes questões de pesquisa:

• RQ1. Quais são as ferramentas de análise, Data Mining e BI mais utilizadas?

• RQ2. Qual a experiência do investigador nessas ferramentas?

• RQ3. Como o uso dessas ferramentas é avaliado por seus clientes?

• RQ4. Quais são os algoritmos de Data Mining mais utilizados?

Para avaliação das questões de pesquisa, foram utilizadas métricas baseadas em frequência, perfazendo o número de respostas por ferramentas utilizadas (RQ1), por níveis de experiência do profissional no uso das ferramentas (RQ2), pelos níveis de utilidade da ferramenta no processo investigativo (RQ3) e pelas técnicas de Data Mining utilizadas (RQ4).

Ainda dentro do planejamento do survey, foi considerada a hipótese de que a maioria das agências de inteligência já utilizavam essas ferramentas no processo de investigação. Desta forma, a hipótese ora testada e que restou elucidada foi:

Hipótese 1

• H0: As unidades que investigam os crimes mais complexos fazem uso de ferramentas de Data Mining e Data Analytics em suas atividades de investigação.

• H1: As unidades que investigam os crimes mais complexos não fazem uso de ferramentas de *Data Mining* e *Data Analytics* em suas atividades de investigação.

As evidências extraídas a partir da pesquisa alertaram para o fato de que 40% dos pesquisados não conheciam processos ETL (*Extract, Transform and Load*) e, dos que conheciam, 15% não utilizavam estas soluções. Mesmo neste cenário de desconhecimento e desuso, 100% dos entrevistados declararam possuir pelo menos uma ferramenta de Data Mining em seu local de trabalho, bem como também declararam (100%) possuir pelo menos uma ferramenta OLAP/BI (*Online Analytical Processing/Business Intelligence*).

Outro ponto importante, que também serviu de suposição de pesquisa e que instiga este estudo, é a evidência de que apenas 2,77% dos pesquisados utilizam diretamente algum algoritmo de Mineração de Dados para extração de conhecimento.

O cenário evidenciado nesta pesquisa inicial aponta que a maior parte dos órgãos especializados em investigação do Brasil ainda não aplica ou desconhece completamente as técnicas de *Data Mining* e de *Data Analytics* em suas atividades investigativas, apoiando a necessidade de realização de pesquisas que ajudem a melhorar este quadro.

Neste contexto, propomos a aplicação e avaliação dos algoritmos de mineração de dados *Expectation Maximization* (EM) e K-Means aos fluxos de transações financeiras de possíveis fraudadores, oriundas de investigações reais, realizadas pelos LAB-LDs de Sergipe. Para isto, tomaremos por base o trabalho de Chen et al. (2014), os quais, em suas conclusões, evidenciaram que o algoritmo EM teve um comportamento melhor que o K-Means, no que concerne à detecção de lavagem de dinheiro. Ainda segundo os autores, foi utilizada apenas uma base de dados de um banco local na Malásia.

Além da limitação da base de dados e da ausência da condução de um processo experimental, “outros estudos precisam ser feitos para determinar o

número ótimo de clusters, de modo que a taxa de detecção falsa seja grandemente reduzida, para evitar que muitas operações normais sejam erroneamente classificadas como anomalias" (Chen et al., 2014). Neste sentido, também poderemos recomendar, ao final deste trabalho, ainda que minimamente, um alinhamento de técnicas, baseadas nos resultados encontrados, como um método a ser utilizado pelos investigadores de Lavagem de Dinheiro em todo o Brasil.

Por fim, esta dissertação vislumbra a condução de um processo experimental, seguindo as diretrizes de (WOHLIN et al., 2012), para também efetuar um comparativo entre os resultados encontrados e as evidências disponíveis na literatura.

Diante de tal cenário, foram elencadas as seguintes questões de pesquisa:

a) Q1: No contexto das análises investigativas conduzidas pelo LAB-LD de Sergipe, o algoritmo EM possui maior eficácia que o K-Means, na detecção de transações financeiras suspeitas?

b) Q2: As eficácias alcançadas pelos algoritmos EM e K-Means, encontradas na literatura (Chen et al., 2014), mantêm-se para o cenário do LAB-LD de Sergipe?

A partir dessas questões, uma hipótese poderá ser formalmente testada:

Hipótese nula H°: O algoritmo de mineração de dados EM, no contexto das transações financeiras sob a análise dos Laboratórios de Tecnologias Contra a Lavagem de Dinheiro de Sergipe, possui a mesma acurácia alcançada pelo algoritmo K-Means.

H 0: $\mu 1$ (Acurácia) = $\mu 2$ (Acurácia)

Hipótese alternativa H^1: O algoritmo de mineração de dados EM, no contexto das transações financeiras sob a análise dos Laboratórios de Tecnologias Contra a Lavagem de Dinheiro de Sergipe, possui acurácia distinta da alcançada pelo algoritmo K-Means.

H 1: μ1 (Acurácia) # μ2 (Acurácia)

Para segunda questão de pesquisa, devido às falhas encontradas no processo de empacotamento dos estudos encontrados, impossibilitando uma replicação completa destes, uma hipótese formal não será testada. No entanto, nosso trabalho possui algumas peculiaridades. Primeiro, o fato de estarmos utilizando como parâmetros, transações financeiras contextualizadas no âmbito dos Laboratórios de Tecnologias Contra a Lavagem de Dinheiro, ou seja, investigações reais. Em segundo lugar, o fato de emularmos o modelo em uma investigação real, forçando, assim, uma melhor compreensão do algoritmo de sucesso. Por fim, estamos trabalhando diretamente com um órgão de investigação de estrutura única no Brasil, o LABLD.

1.3 Justificativa

Este projeto objetiva realizar uma avaliação experimental que permita mensurar as eficácias dos algoritmos de aglomeração *Expectation Maximization (EM)* e *K-Means*, quando aplicados a transações financeiras, oriundas de um caso real, investigado pelo Laboratório de Tecnologia contra a Lavagem de Dinheiro em Sergipe. Posteriormente, os resultados encontrados serão comparados com resultados de trabalhos similares publicados na literatura. Isto justificará o esforço a ser depreendido, pois diminuirá a lacuna evidenciada no *survey* apresentado nesta pesquisa, gerando uma produção científica que poderá recomendar um método a ser utilizado pelos investigadores de Lavagem de Dinheiro em todo o Brasil, bem com alavancar novas pesquisas que incrementem o uso de algoritmos de mineração de dados em órgãos que realizam investigações complexas, a exemplo dos órgãos de inteligência.

1.4 Objetivos da pesquisa

O objetivo deste projeto foi avaliar, por meio de um processo experimental, a eficácia dos algoritmos EM e K-Means sobre bases de dados de transações financeiras investigadas pelos Laboratórios de Tecnologia de Combate à Lavagem de Dinheiro de

Sergipe, comparando as evidências encontradas com os resultados publicados na literatura.

1.4.1 Objetivos Específicos

Para que pudéssemos lograr êxito na realização do objetivo geral, podemos enumerar os seguintes objetivos específicos:

- Construir e aplicar um Survey, para apresentar um levantamento acerca da utilização de ferramentas de *Data Analytics* e de *Data Mining* nos principais órgãos brasileiros de combate ao crime organizado;
- Fazer um Mapeamento Sistemático com a finalidade de identificar pesquisas sobre a aplicação de Data Mining no combate à Lavagem de Dinheiro;
- Realizar experimento para avaliar as eficácias dos algoritmos EM e K-Means, no contexto dos LABLDs de Sergipe, bem como comparar os resultados encontrados com os resultados publicados na literatura.

1.5 Metodologia

A metodologia adotada para o trabalho envolveu, inicialmente, um Survey realizado no Brasil, aplicado pelo grupo de pesquisa do qual o autor desta dissertação faz parte, o qual foi publicado e apresentado por este (Santos; Mangueira; Oliveira & Colaço Júnior, 2017), descrevendo um levantamento acerca da utilização de Data Mining e Data Analytics por órgãos de investigação. Ato contínuo, foi feito um mapeamento sistemático (KITCHENHAM, 2004), tendo por finalidade encontrar pesquisas sobre algoritmos de Data Mining aplicados à Segurança e à detecção de Lavagem de Dinheiro.

O mapeamento evidenciou um grande número de trabalhos que utilizaram agrupamento para o combate à LD, perfazendo 14 trabalhos e tendo como destaque o algoritmo K-means. Este cenário motivou a busca por um trabalho científico específico que tivesse lidado apenas com transações financeiras e com o algoritmo K-means. Desta forma e neste contexto, o trabalho de Chen et al. (2014) foi selecionado por apresentar

uma base de dados com alta similaridade à base de dados analisada pelos LABLDs do Brasil, comparando a eficácia dos algoritmos K-means e EM, para detecção de transações suspeitas. Em suma, para que fosse possível a combinação de evidências experimentais, este trabalho, além de avaliar o K-means, também avaliou o algoritmo EM. A seguir, esta avaliação será enquadrada quanto a sua taxionomia.

Para averiguar as possibilidades de Data Analytics para segurança pública e combate à LD em Sergipe, foi adotado o método científico hipotético-dedutivo que, segundo Dresch, Lacerda e Antunes Júnior (2015), a partir de conhecimentos prévios, permite identificar um problema, estabelecer e testar hipóteses que podem resultar em previsões e explicações a cerca de um fenômeno. Neste contexto, as avaliações dos modelos de conhecimento (K-means e EM) foram feitas a partir de um experimento controlado.

No que diz respeito à classificação desta pesquisa, podemos citar, quanto à natureza, como sendo aplicada, pois produziu conhecimento para aplicação de seus resultados com o objetivo de contribuir para fins práticos, visando à solução imediata do problema encontrado na realidade (Appolinário, 2007; Dresch; Lacerda; Antunes Júnior, 2015). Quanto à abordagem dos dados, foi considerada quantitativa, pois as variáveis estão associadas a valores numéricos, foram obtidas de medições objetivas e analisadas estatisticamente (Pimentel; Fuks, 2012).

Quanto aos objetivos, podemos afirmar como sendo explanatória e explicativa, pois objetiva identificar os fatores primordiais para a ocorrência de um fenômeno, bem como é estabelecida uma hipótese de causa-e-efeito sobre o fenômeno estudado (Pimentel; Fuks, 2012). Também será classificada como experimental, pois, segundo Wohlin et al. (2012), um experimento é um estudo empírico que manipula um fator ou variável de um ambiente controlado. Assim, seguindo a linha experimental do grupo de pesquisa do autor desta dissertação, com origens em Maryland e fundamentada nos trabalhos de Victor Basili (Basili et al., 2014), pai da Engenharia de Software Experimental, foi realizada a seleção de conjuntos de dados, para análise de fatores, com tratamentos representados pela utilização de modelos de conhecimento num ambiente controlado real (In Vitro e In Vivo) e, finalmente, uma avaliação do aumento da eficácia em investigações realizadas.

O capítulo 5, capítulo que descreve o experimento, é autocontido, do ponto de vista metodológico, pois descreve, em cada seção, os passos adotados no processo experimental realizado.

1.6 Organização da proposta

Este documento está organizado de acordo com a Instrução Normativa Nº 02/2015/PROCC, a qual permite que a Dissertação seja "uma compilação de artigos científicos submetidos ou publicados em veículos com Qualis, desde que seja contextualizada com seções de Introdução e Conclusão, não limitada a estas". São 6 capítulos que fornecem uma base conceitual e experimental para o entendimento sistêmico. Os tópicos a seguir descrevem o conteúdo de cada um dos capítulos:

- O Capítulo 1 apresenta esta Introdução, explicando as justificativas juntamente com as hipóteses levantadas;
- O Capítulo 2 traz um breve Referencial teórico acerca da temática abordada;
- O Capítulo 3 traz um artigo que foi apresentado e publicado no XIII Simpósio Brasileiro de Sistemas de Informação - Lavras/MG - SBSI 2017, Qualis B2;
- O Capítulo 4 traz um artigo aceito no *16th International Conference on Information Technology : New Generations - ITNG 2019*, Qualis B1, a ser apresentado em Las Vegas;
- No capítulo 5, são descritos o Planejamento, Operação e Resultados do Experimento;
- Finalmente, no capítulo 6, é apresentado um compilado de conclusões, contribuições e sugestões de trabalhos futuros.

CAPÍTULO 2

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo descreve uma visão geral dos principais conceitos pertinentes à pesquisa, tendo em vista dar um embasamento teórico para a mesma. Deste modo, aqui são definidos Lavagem de Dinheiro e os algoritmos de aglomeração Expectation Maximization (EM) e K-Means, os quais serão adotados para execução deste projeto.

2.1 Lavagem de Dinheiro

De fato, várias são as terminologias ou palavras para definir o que é Lavagem de Dinheiro (LD). No Brasil, o caput do Art 1° da lei n° 12.683, de 2012, traz: "Ocultar ou dissimular a natureza, origem, localização, disposição, movimentação ou propriedade de bens, direitos ou valores provenientes, direta ou indiretamente, de infração penal. "

Na literatura, alguns autores se referem como "a atividade ou processo que lida com ativos vindos das atividades criminosas, com o objetivo de disfarçar sua origem ilícita e fazer com que pareçam legítimos " (Zhang et al., 2003). LD é considerado um crime vultoso em criminologia, sendo identificado como um dos mais relevantes na sociedade atual (Zhang et al., 2003), além de ser, frequentemente, um crime de transição que ocorre em íntima relação com outros crimes, tais como tráfico de drogas, terrorismo e tráfico de armas (Schott, 2006).

Criminosos da sociedade atual, sociedade esta baseada no uso e manejo tecnológico, valem-se de todos os meios disponíveis para lavar os lucros de suas atividades ilegais. Em resposta, a comunidade internacional tem buscado esforços contra lavagem de dinheiro (GAO et al., 2009). Normalmente, instituições financeiras usam processos semiautomatizados para marcar transações suspeitas de LD, baseados em médias e irregularidades pré-determinadas (Alexandre et al., 2016). Sistemas contra lavagem de dinheiro são primordiais e fundamentais para auxiliar os órgãos de governo e os órgãos de controle da atividade financeira no país, os quais lutam contra esta prática.

2.2 Expectation Maximization - EM

O Algoritmo de *Expectation Maximization* decorre de uma série de técnicas estatístico estocásticas cuja finalidade é estimar parâmetros em decorrência de variáveis escondidas dentro de um conjunto de amostras. A modelagem estatística, aqui posta, refere-se às distribuições de probabilidades que, notadamente, adotam parâmetros estocásticos em detrimento de determinísticos. Desta forma, o algoritmo EM consegue processar tanto variáveis ocultas quanto os parâmetros pré-definidos.

Segundo Dempster et al. (1977), o algoritmo maximização da expectativa é descrito como "uma ferramenta computacional utilizada para o cálculo do estimador de máxima verossimilhança (EMV) de forma iterativa, e é principalmente utilizado em problemas envolvendo dados incompletos".

Para Paula (2013), o algoritmo de maximização da expectativa (ou EM, do inglês *Expectation Maximization*) é um procedimento iterativo para maximização de $L(\Theta)$.

Vitor (2011), apud Casella e Berger (2010), definem o algoritmo EM como um algoritmo que seguramente converge para o estimador de máxima verossimilhança – EMV - e tem como base a ideia de substituir uma difícil maximização da verossimilhança por uma sequência de maximizações mais fáceis.

2.3 K-Means

O algoritmo *K-means* é uma técnica computacional baseada em clusterização que visa particionar o número de observações entre k grupos, aproximando-se ao máximo da média e visando minimizar o erro do agrupamento.

Segundo Wang et Bai (2016) apud Celebi et al. (2013); Peña et al. (1999); Celebi & Kingravi (2012, 2014), apesar da popularidade, o algoritmo k-means é sensível à escolha das condições de iniciais de partida.

Ainda segundo Wang et Bai (2016), embora o algoritmo de *k-means* global seja um metodo deterministico e muitas vezes seja bem-sucedido, às vezes, o novo centro de

cluster pode ser um outlier. Desta forma, podem surgir clusters com um único ponto, sendo necessária a análise de como tratar estes casos.

Um ponto positivo no uso do algoritmo k-means é o fato deste apresentar uma eficiência na detecção de intrusão (Eslamnezhad et Varjani, 2014).

De fato, a análise de cluster é um ponto muito importante no campo da mineração de dados, sendo possivel fracionar um conjunto de dados em várias classes ou clusters, de modo a ter um alto grau de semelhança entre os objetos de dados no mesmo cluster e um baixo grau de similaridade entre os objetos de dados em grupos diferentes.

2.4 Matriz de Confusão

A matriz de confusão é uma ferramenta bastante útil para analisar a capacidade de predição de um classificador em determinar a classe de vários registros (Han et al., 2011).

Dadas as classes m, uma matriz de confusão é uma tabela de dimensão m x m, onde, para cada classificação possível, existe uma linha e coluna correspondente, ou seja, os valores das classificações serão distribuídos na matriz de acordo com os resultados, assim gerando a matriz de confusão para as classificações realizadas. As linhas correspondem às classificações corretas e as colunas representam as classificações realizadas pelo classificador (Bramer, 2007).

Quando existem apenas duas classes, uma é considerada como positive (em nosso contexto, “Suspeita”) e a outra como negative (“Normal”) (Bramer, 2007). Assim, podemos ter quatro resultados possíveis:

- *True Positive* (*TP*): uma instância de classe *positive* (uma evidência) é classificada corretamente como *positive* (“Suspeita”);
- *False Negative* (*FN*): uma instância de classe *positive* é classificada incorretamente como *negative* (“Normal”);
- *True Negative* (*TN*): uma instância de classe *negative* pode ser classificada corretamente como *negative*;
- *False Positive (FP): uma instância de classe negative é classificada incorretamente como positive.*

As predições corretas e errôneas para as duas classes podem ser dispostas em uma única matriz, conforme é vista na Tabela 1 (Bramer, 2007).

Tabela 1. Matriz de Confusão

Actual Class	Predicted Class	
	Positive	*Negative*
Positive	True Positive (TP)	False Negative (FN)
Negative	False Negative (FN)	True Positive (TP)

2.5 Métricas de Qualidade

Com a matriz de confusão apresentada na seção 2.4, podemos utilizar como principal métrica de qualidade, a acurácia.

2.5.1 Acurácia

É o percentual de instâncias classificadas corretamente (ver Figura 1).

Figura 1. Fórmula da Acurácia

$$\text{acurácia} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

2.5.2 Log Score

Log Score é baseado na métrica de pontuação ao local apropriada (Abramovici et al., 2008), é o logaritmo da probabilidade real para cada caso, somado, e depois dividido pelo número de linhas no conjunto de dados de entrada. Como a probabilidade é representada como uma fração decimal, as pontuações de log são sempre números negativos. Um número mais próximo de 0 é uma pontuação melhor. Visto que contagens brutas podem ter distribuições muito irregulares ou distorcidas, uma contagem de log é semelhante a uma porcentagem.

2.5.3 Root Mean Square Error (RMSE)

Raiz do erro quadrático médio é uma medida de acurácia, para comparar erros de previsão de modelos diferentes em um conjunto de dados específicos e nao entre conjuntos de dados, pois é dependente da escala (Hyndman et al., 2006). Como o nome já elucida, é a raiz quadrada do erro médio para todos os casos, dividida pelo número de casos.

O *Root Mean Square Error* (RMSE) também é um avaliador popular para modelos preditivos. A contagem calcula a média dos resíduos para cada caso, produzindo um unico indicador de erro para o modelo.

CAPÍTULO 3

3 ARTIGO I - SURVEY

Um Survey sobre a utilização de técnicas de Data Mining e Data Analytics por agências de investigação criminal do Brasil

Alternative Title: A Survey on the use of Data Mining and Data Analytics techniques by Brazilian criminal investigation agencies

Rafael Meneses Santos[a], Fábio Mangueira da Cruz Nunes[a,b], Manoela dos Reis Oliveira[a], Methanias Colaço Júnior [a,b,c]

[a]Postgraduate Program in Computer Science - PROCC.
UFS – Federal University of Sergipe
São Cristóvão/SE - Brasil.
rafaelsantos@ufs.br, fabio.dipolcgi@gmail.com, manoelareisoliveira@gmail.com, mjrse@hotmail.com

[b]Department of Public Security of Sergipe – SSP/SE
[c]Competitive Intelligence Research and Practice Group – NUPIC
Information Systems Departament - DSI
UFS – Federal University of Sergipe
Itabaiana/SE - Brasil

RESUMO

Em investigações criminais complexas, os envolvidos lidam com uma quantidade enorme e complexa de dados que necessitam de recursos computacionais especializados na extração de informações e correlações relevantes para o processo investigativo. Neste cenário, é necessário que haja apoio computacional, desde a etapa de armazenamento e integração entre diferentes bases de dados, até a etapa de análise estatística e descoberta de padrões. Este artigo discute os resultados de um Survey aplicado aos principais órgãos de combate ao crime organizado, tais como as agências de Inteligência de Segurança Pública – ISP, os Laboratórios de Tecnologia de Combate à Lavagem de Dinheiro – LABLDs e os Grupos de Atuação Especial de Repressão ao Crime Organizado – GAECO. O objetivo principal foi o de conhecer o cenário atual da utilização de ferramentas de análise de dados nessas agências, projetando as necessidades de pesquisa e investimentos nesta área. Entre os resultados encontrados, observou-se que 40% dos pesquisados não conhecem e 15% não utilizam soluções de ETL (Extract, Transform and Load), apesar de todos (100%) declararem possuir pelo menos uma ferramenta de Data Mining no seu local de trabalho, bem como também declararem (100%) possuir pelo

menos uma ferramenta de OLAP/BI (Online Analytical Processing/Business Intelligence). Por fim e com proeminente destaque, apenas 2,77% dos pesquisados utilizam diretamente algum algoritmo de Mineração de Dados para extração de conhecimento. Este cenário evidencia, inicialmente, que a maior parte dos órgãos especializados em investigação do Brasil ainda não aplica efetivamente as técnicas de Data Mining e de Data Analytics em suas atividades.

SBSI 2017, June 5th–8th, 2017, Lavras, Minas Gerais, Brazil.

Palavras-Chave

Inteligência de Segurança Pública (ISP), Investigação, Segurança Pública, *Data Mining*, *Data Analytics*.

ABSTRACT

In complex criminal investigations, those involved deal with a huge and complex amount of data that requires computational resources specialized in extracting information and correlations relevant to the investigative process. In this scenario, it is necessary to have computational support, from storage and integration between different databases, to statistical analysis and pattern discovery. This article discusses the results of a survey applied to the main organs to combat organized crime, such as Public Security Intelligence agencies - ISP, Anti-Money Laundering Laboratories - LABLD and the Special Action Groups on Organized Crime Repression - GAECO. The main objective was to know the current scenario of the use of data analysis tools in these agencies, projecting research and investments needs in this area. Among the results found, 40% of respondents did not know and 15% did not use ETL solutions, although all (100%) declare to have at least one Data Mining tool in your workplace, as well as declaring (100%) to have at least one OLAP / BI tool. Finally, the results highlighted that only 2.77% of respondents directly use some Data Mining algorithm for knowledge extraction. This scenario shows, initially, that most of Brazil's specialized investigation agencies do not yet effectively apply Data Mining and Data Analytics techniques in their activities.

CCS Concepts

• **Information systems→ Database design and models; Information integration; Decision support systems; Data mining.**

Keywords

Public Security Intelligence (ISP), Investigation, Public Security, Data Mining, Data Analytics.

3.1 Introdução

No Brasil, os órgãos especializados em investigações complexas têm o papel fundamental de produzir conhecimento e informações de valor significativo, podendo assim subsidiar as investigações com grande volume de dados e trazer luz aos fatos contidos nos afastamentos de sigilos quebrados por decisão judicial, além dos que estão disponibilizados na web e que atendam às necessidades investigativas dos processos [1]. Contudo, a atividade de ISP merece um destaque especial, pois deverá exercer sempre sua função básica de produzir conhecimento de interesse e utilidade para a instituição policial ou para outro órgão de controle que também exerça atividade correlata à investigação policial.

Desta forma, a legislação pátria define formalmente a atividade de inteligência como "a atividade que objetiva a obtenção, análise e disseminação de conhecimentos dentro e fora do território nacional, sobre fatos e situações de imediata ou potencial influência sobre o processo decisório e a ação governamental, e sobre a salvaguarda e a segurança da sociedade e do Estado" [2].

Nesta corrente, a Doutrina Nacional de Inteligência de Segurança Pública também aporta, definindo a ISP como "o exercício permanente e sistemático de ações especializadas para identificar, avaliar e acompanhar ameaças reais ou potenciais na esfera de Segurança Pública, basicamente orientadas para produção e salvaguarda de conhecimentos necessários para subsidiar os tomadores de decisão, para o planejamento e execução de uma política de Segurança Pública e das ações para prever, prevenir, neutralizar e reprimir atos criminosos de qualquer natureza que atentem à ordem pública, à incolumidade das pessoas e do patrimônio" [2].

Baseados nesta Doutrina, quando nos deparamos com um universo de ações criminosas que são executadas diariamente em nossa sociedade, para as quais o nosso mecanismo de defesa e proteção é a segurança pública e demais órgãos de controle, percebemos que se faz relevante entender o seu meio de funcionamento, sua estrutura e os seus resultados, bem com se seus esforços têm usado técnicas automatizadas para analisar diferentes tipos de crimes, mesmo sem um arcabouço unificador que descreva como aplicá-las [3].

Dentro do campo da atividade investigativa, a investigação, na prática, nada mais é do que uma busca constante por um grande volume de dados, de maneira a levar o investigador, no sentido amplo da palavra, à identificação e ao evidenciamento de informações que possam trazer à luz dados concretos sobre o fato investigado. Mais especificamente, é a busca por informações que possam levar os investigados aos rigores da lei.

Essa realidade prática impõe uma dependência das informações e dos dados disponíveis em diversas bases, sejam estas de fontes primárias ou secundárias, perfazendo um fluxo de informações que necessita de um trabalho diferenciado, a ser exercido por profissionais capazes de produzir os resultados esperados.

Diante dessa problemática existente nas atividades policiais, um modelo de bases de dados ideal deve abranger dados dos órgãos estaduais e federais, para que haja efetividade na busca por resultados e soluções. Este modelo exigirá um conjunto de ações que possibilitem a sistematização do fluxo de dados e desburocratização em torno da comunicação entre as agências, de modo que as informações trafeguem numa constância ininterrupta e contribuam para a solução de crimes.

Desta forma, fica notória a necessidade de integração de dados para realização de um trabalho eficiente, sem ser desconsiderada a complexidade desta atividade, uma vez que os dados precisam ser modelados e os metadados padronizados, além da exigência de um ambiente específico para processamento e análise, o qual está diretamente relacionado com a qualidade dos dados que são armazenados numa base histórica [4, 11].

Essa base histórica tende a ser volumosa e complexa, com características do fenômeno Big Data [6], as quais posicionam o processo investigativo como desafiador e ávido por técnicas que auxiliem as suas atividades. Neste sentido, para assessorar na extração de conhecimento, técnicas de Data Mining (Mineração de Dados) e Data Analytics (Análise de Dados) são abordagens muito utilizadas para descobrir padrões e extrair informações que podem ser úteis a tarefas de investigação [5].

Diante desta realidade, faz-se necessária uma avaliação, em escala nacional, para entender como as agências de inteligência e demais órgãos de investigação e controle no Brasil estão utilizando os recursos computacionais para o tratamento, armazenamento e análise dos dados utilizados nas atividades de investigação.

Neste artigo, apresentamos um Survey com agências de ISP do Brasil, tais como os Laboratórios de Tecnologia de Combate à Lavagem de Dinheiro e outros órgãos de controle que exerçam atividades investigativas. Foram levantados dados relacionados ao tratamento, armazenamento, integração e análise de informações, com o foco direcionado para a utilização de ferramentas e técnicas de Data Mining e Data Analytics. Para tanto, foi utilizado um questionário que consiste em 21 questões e divide-se em 5 grupos. O primeiro grupo caracteriza o entrevistado. Os outros quatro grupos são divididos em questões sobre armazenamento e preparação dos dados, Data Mining, análise estatística inferencial (Data Analytics) e a utilização destes recursos pela agência.

O questionário foi disponibilizado na internet, por meio da ferramenta SurveyMonkey, e pessoas que trabalham nas agências de inteligência foram convidadas a respondê-lo. Foi obtida uma amostra de 108 respostas, com representantes de todos os estados da federação, durante os dias 05/08/2016 e 17/10/2016.

Assim, observou-se que 44% dos pesquisados não conhecem e 15% não utilizam soluções de ETL, apesar de todos os pesquisados (100%) declararem possuir pelo menos uma ferramenta de Data Mining no seu local de trabalho e pelo menos uma ferramenta de OLAP/BI (Business Intelligence). Nesta linha, apenas 2,77% dos pesquisados utilizam diretamente algum algoritmo de Mineração de Dados para extração de conhecimento, o que nos permite inferir, inicialmente, que a maior parte destes órgãos especializados do Brasil ainda não aplica efetivamente as técnicas de Data Mining e Data Analytics em suas atividades investigativas.

O restante do trabalho está estruturado como segue. A Seção 2 apresenta uma discussão sobre os trabalhos relacionados. Na Seção 3, é abordado o objetivo do Survey, passando pela seleção de participantes, instrumentação e operação, até a análise e interpretação dos resultados colhidos. A Seção 4 apresenta as ameaças à validade da pesquisa de campo. Por fim, a Seção 5 apresenta as conclusões que puderam ser extraídas desse estudo, bem como sugestões de possíveis trabalhos futuros.

3.2 Trabalhos Relacionados

Não foram encontrados Surveys científicos com o mesmo objeto de pesquisa deste artigo, inclusive tratando da utilização de técnicas de Data Mining e Data Analytics nas

agências de investigação no Brasil. Este fato aumenta a importância dos dados aqui apresentados. Além disso, nosso trabalho difere de grande parte dos Surveys apresentados na área de computação, pois, do ponto de vista das agências e dos laboratórios, não se trata de amostragem, foi realizado um censo, considerando profissionais entrevistados nos 27 estados, com respostas de agências de inteligência de todo o país, bem como de todos os laboratórios de tecnologia de combate à lavagem de dinheiro.

O uso de Data Mining e ferramentas de BI na Segurança Pública é alvo de algumas pesquisas no Brasil [12, 13, 14]. Braz, Coan e Rosseti [12] desenvolveram um sistema baseado em Data Mining para análise de dados georeferenciados, permitindo melhor entendimento sobre ocorrências armazenadas na base de dados da Polícia Militar. Dados georeferenciados podem oferecer uma análise ainda mais precisa de ocorrências, dando oportunidade à aplicação de técnicas de reconhecimento de padrão [13].

Leite et al. [14] propuseram uma ferramenta para melhorar a visualização de dados públicos da Segurança Pública, por meio de ferramentas OLAP e Data Marts.

3.3 Survey

Esta seção apresenta todas as etapas referentes à realização do Survey, desde seu objetivo, passando pela seleção de participantes, instrumentação, operação, até a análise e interpretação das respostas coletadas.

3.3.1 Objetivo

O objetivo geral deste Survey é mapear a utilização de técnicas e ferramentas de armazenamento, integração, Data Mining e Data Analytics, no processo de investigação que transcorre em agências de inteligência brasileiras, laboratórios de tecnologia de combate à lavagem de dinheiro e demais unidades de igual monta que utilizam o arcabouço computacional aqui pesquisado. Este objetivo é formalizado usando parte do modelo GQM proposto por Basili e Weiss [7], como apresentado por Van Solingen e Berghout [8]: Analisar as atividades de investigação criminal, com o propósito de caracterizar, com respeito à utilização de ferramentas de armazenamento, integração, Data Mining e Data Analytics, do ponto de vista de investigadores e cientistas de dados, no contexto de agências governamentais brasileiras que exercem atividades investigativas. Baseadas neste objetivo, foram formuladas as seguintes questões de

pesquisa:

- RQ1. Quais são as ferramentas de análise, *Data Mining* e BI mais utilizadas?
- RQ2. Qual a experiência do investigador nessas ferramentas?
- RQ3. Como o uso dessas ferramentas é avaliado por seus clientes?
- RQ4. Quais são os algoritmos de Data Mining mais utilizados?

Essas questões de pesquisa foram utilizadas para derivar as perguntas do questionário, analisadas nas próximas seções.

3.3.2 Planejamento

3.3.2.1 Formulação de Hipóteses

Para avaliar as questões de pesquisa, serão utilizadas métricas baseadas em frequência, perfazendo o número de respostas por ferramentas utilizadas (RQ1), por níveis de experiência do profissional no uso das ferramentas (RQ2), pelos níveis de utilidade da ferramenta no processo investigativo (RQ3) e pelas técnicas de Data Mining utilizadas (RQ4).

Tendo o objetivo e métricas definidas, será ainda considerada a hipótese de que, atualmente, a maioria das agências de inteligência já utiliza essas ferramentas no processo de investigação. Desta forma, a hipótese que queremos testar é:

H0: As unidades que investigam os crimes mais complexos fazem uso de ferramentas de Data Mining e Data Analytics em suas atividades de investigação.

H1: As unidades que investigam os crimes mais complexos não fazem uso de ferramentas de Data Mining e Data Analytics em suas atividades de investigação.

3.3.2.2 Seleção de Participantes e Amostra

A seleção dos participantes ocorreu por censo, se considerarmos as agências de inteligência e os Laboratórios de Tecnologia de combate à Lavagem de Dinheiro (LABLDs). Em todos os 27 (vinte e sete) estados da federação, foram consultados órgãos de inteligência e os LABLDs supracitados, podendo estes ser ou não organizacionalmente vinculados ao órgão de inteligência do estado em que atuam.

Desta forma, buscou-se obter informações em todos os órgãos existentes, sejam eles atrelados aos órgãos de segurança pública ou não, pois sabemos que não somente os

órgãos de segurança pública atuam no combate aos crimes mais complexos ou de natureza organizacional mais elaborada. Os Ministérios Públicos também têm exercido um papel preponderante e fundamental neste contexto, principalmente no que tange à utilização dos LABLDs como órgãos de apoio, assessoramento e processamento das informações. Esta relação de suporte faz destes laboratórios uma parte integrante de quase todas as Polícias Judiciárias e Ministérios Públicos Estaduais.

3.3.2.3 Metodologia

Foi projetada a execução de um piloto com os profissionais que tivessem uma relação direta com análise de dados e que utilizassem extrações de informação e de conhecimento para apoio ao processo de investigação criminal.

A amostra para o piloto deve ser menor, com fins de identificar possíveis problemas e inconsistências nas perguntas. Esse pré-teste é necessário e visa melhorar o instrumento da pesquisa, sendo executado da mesma forma como será aplicado. A seleção de quem irá participar do pré-teste é flexível, entretanto, recomenda-se que as pessoas sejam capacitadas para responder as perguntas.

Por fim, foi planejado o contato com os órgãos, solicitando a indicação de um analista com conhecimento ou atuação no setor responsável pela tecnologia da informação, como também outro ator que utilizasse uma das tecnologias abordadas, dentro de cada órgão respondente. A atuação pôde ser no nível decisório ou apenas de assessoramento.

Desta forma, a meta foi atingir todos os atores envolvidos diretamente com a atividade investigativa de alta complexidade, obtendo respostas das Polícias Judiciárias Estaduais, como também dos Ministérios Públicos Estaduais, com um mínimo de um ator por órgão.

3.3.2.4 Instrumentação

O questionário foi desenvolvido na ferramenta especialista SurveyMonkey [13] e distribuído por meio da internet. Contém uma apresentação inicial, seguida das perguntas referentes à utilização de ferramentas de armazenamento, integração, Data Mining e Data Analytics.

3.3.3 Operação

3.3.3.1 Aplicação

Nesta etapa, acontece a efetiva realização da pesquisa. Tudo que foi planejado nas etapas anteriores passa agora a concretizar-se.

Inicialmente, um piloto do questionário foi aplicado a três agentes de inteligência da Divisão de Inteligência e Planejamento Policial (DIPOL), órgão da polícia civil de Sergipe, a dois agentes da Gerência de Inteligência (GI) do Comando de Operações Especiais (COE) da polícia militar do estado de Sergipe, assim como a dois agentes do Laboratório de Tecnologia de Combate à Lavagem de Dinheiro (LABLD), também da polícia civil do mesmo estado. Estes profissionais foram definidos na metodologia e selecionados por julgamento, sem participação no Survey final, mas com contribuição para modificações, tornando o questionário mais claro e objetivo.

Em seguida, foram contatados todos os LABs do país, sejam de Polícia Judiciária ou do Ministério Público Estadual, bem como 27 órgãos de inteligência das agências estaduais, para que indicassem os profissionais que laborassem essencialmente com tecnologia da informação e que atuassem com análise e tratamento de dados. Os indicados foram convidados a responder ao questionário, o qual foi encaminhado via url *(Uniform Resource Locator).*

3.3.3.2 Coleta e Validação de Dados

Mesmo tendo sido utilizada uma ferramenta especialista para a construção de Survey, o SurveyMonkey2, foi verificado se os resultados eram realmente coerentes com os apontados pela mesma, assim como o total de respostas. Além disso, como formas de validação, foram averiguados os e-mails dos participantes, os estados a que eles pertenciam, bem como confirmada a existência da agência apontada.

[2] https://pt .www.surveymonkey.com

3.3.4 Análise e Interpretação dos Dados

3.3.4.1 Dados sobre Perfil e Infraestrutura Básica

Após apresentação do Survey ao público, os participantes começaram a enviar as respostas via SurveyMonkey. O Survey foi respondido por participantes de todos os estados, incluindo todos os LABLDs, podendo estes estar ou não vinculados a algum núcleo, departamento, superintendência de inteligência ou qualquer outro órgão que exerça a atividade de inteligência, quer seja no âmbito do Ministério Público ou nas Secretarias de Segurança Pública, ou ainda em Secretarias de Defesa Social, a exemplo do estado de Pernambuco.

Inicialmente, algumas questões foram elaboradas com a finalidade de identificar o perfil dos entrevistados. Na Figura 1, é apresentado um gráfico referente ao órgão de origem de trabalho dos participantes. É possível verificar maior incidência de participação da Polícia Civil (43,52%), seguida pela Polícia Militar (22,15%) e o Ministério Público Estadual (22,22%).

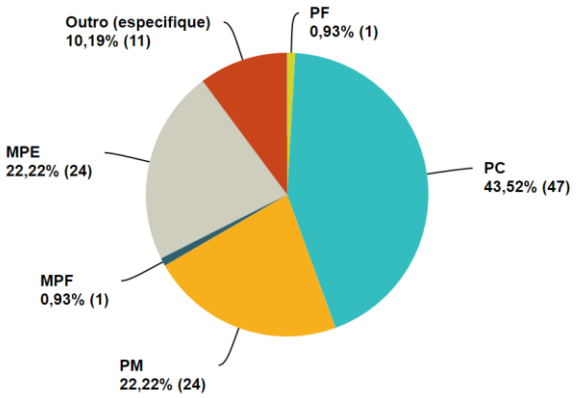


Figura 1. Respostas sobre órgão de trabalho.

Ainda com relação ao perfil dos respondentes, avaliou-se o que diz respeito ao cargo que ocupa no órgão que exerce suas funções. Neste quesito, como apresentado pelo gráfico representado na Figura 2, 26,85% dos participantes ocupam o cargo de Agente de Polícia, 15,74% são Analistas, 11,11% são Delegados, 6,48% são Escrivães de Polícia, 9,26% são Policiais Militares, conhecidos como praças, e 7,41% Oficiais de Policia Militar. Além disso, os pesquisados foram questionados sobre a área em que atuam. Foram classificadas quatro grandes áreas de investigação, considerando a condição e a

capacidade de abordagem dos crimes mais complexos, os quais, notadamente, pela própria natureza, costumam utilizar recursos computacionais que fazem parte do Survey. As respostas, como vistas na Figura 3, foram: Inteligência (62,04%), Combate à Lavagem de Dinheiro (46,30%), Estatísticas e Análise Criminal (9,26%) e Crimes Cibernéticos (5,56%).

No primeiro grupo de perguntas, tratando sobre armazenamento e integração de dados, os entrevistados foram questionados sobre quais os bancos de dados usados. Neste contexto, 48,75% dos órgãos usam o SGBD Oracle e 42,50% usam o Microsoft SQL Server. As respostas apontam uma predominância de softwares pagos no quesito de armazenamento de dados. Softwares de código aberto como MySQL e PostgreSQL são usados por 25,00% e 7,50% dos entrevistados, respectivamente. Uma boa parte (28,75%) dos entrevistados não conhece o SGBD utilizado. Os dados completos são apresentados no gráfico da Figura 4.

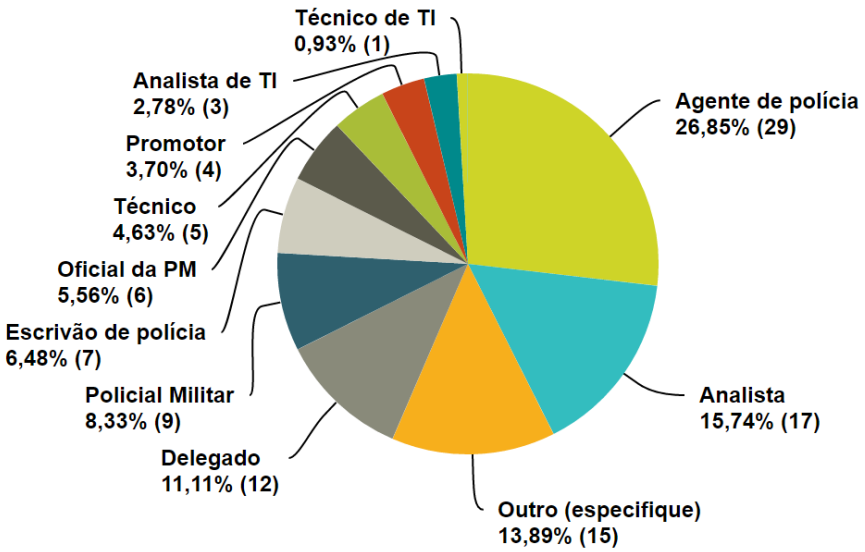


Figura 2. Respostas sobre cargo do pesquisado.

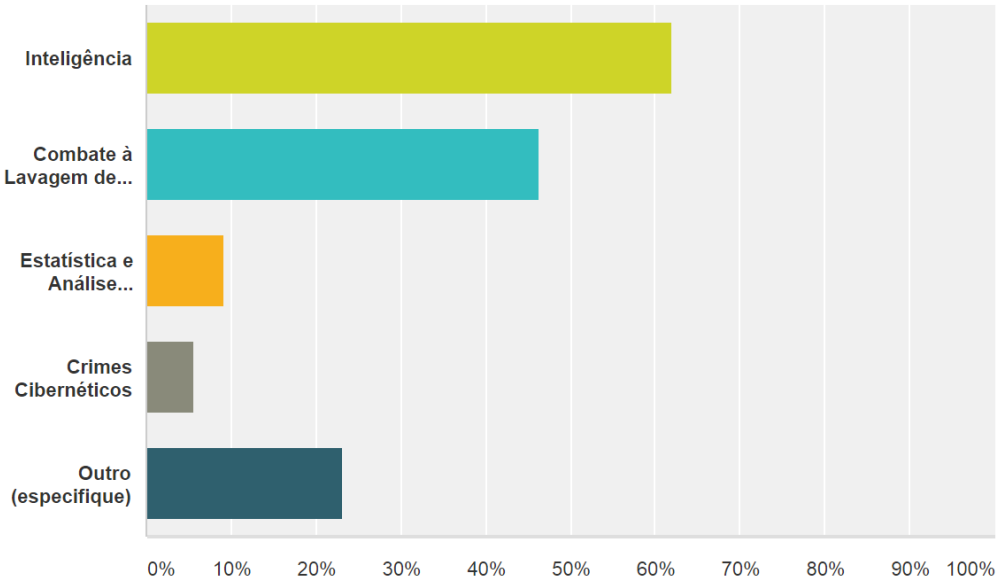


Figura 3. Respostas sobre a área de atuação.

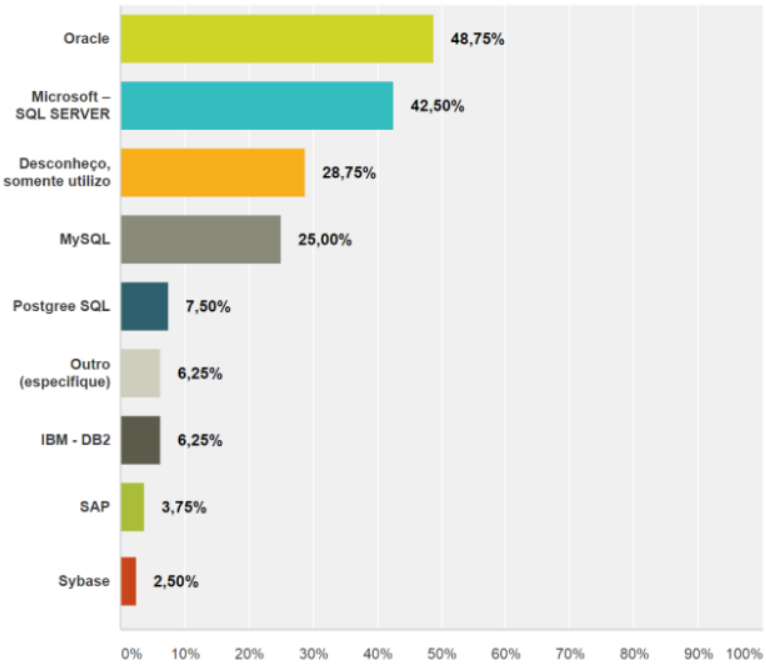


Figura 4. Respostas sobre banco de dados.

Ato contínuo, o Survey apresenta perguntas sobre as ferramentas e técnicas usadas em diversas etapas do tratamento da informação.

Sobre a integração de dados através de processo ETL, 40,00% dos respondentes não conheciam o procedimento e 15,00%, muito embora pudessem conhecer o processo de ETL, não o utilizava. Os órgãos que utilizam tal procedimento afirmam usar ferramentas da Microsoft (27,50%), IBM (16,25%), SAS (15,00%) e Oracle (10,00%).

No último grupo de perguntas, que diz respeito ao auxílio das ferramentas nas atividades de investigação, foi questionado aos pesquisados como eles avaliam o nível de conhecimento das ferramentas escolhidas. Neste quesito, como apresentado pelo gráfico representado na Figura 5, apenas 28,17% avaliaram o seu conhecimento como bom, a maioria (40,85%) avalia como regular e 30,99% avaliam o seu conhecimento como ruim ou péssimo. Esta alta porcentagem de autoavaliações negativas pode indicar falta de treinamento e incentivo ao uso das ferramentas.

A figura 6 representa um gráfico de respostas sobre a avaliação da utilidade e aplicabilidade das ferramentas na atividade de investigação, sendo essa a última questão sobre o auxílio das ferramentas nas atividades de investigação.

Como pode ser observado nas repostas, quase metade dos pesquisados (47,89%) classifica como boas e 11,27% classificam como excelente, porém 18,31% classificam como ruins ou péssimas a aplicabilidade e utilidade dessas ferramentas. Isto pode, em colaboração com a questão anterior, reforçar a ideia de que os pesquisados não estão sendo treinados adequadamente no uso das ferramentas.

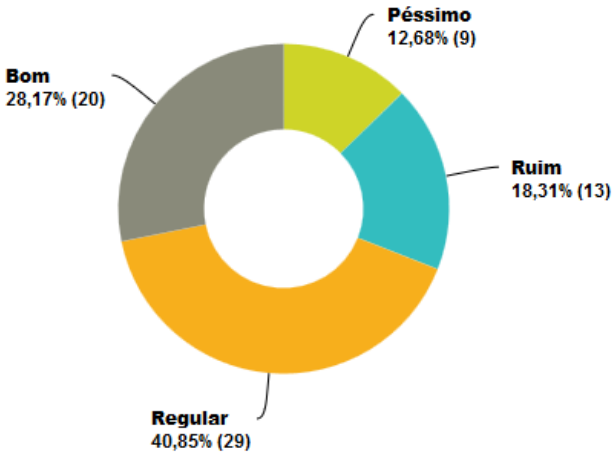


Figura 5. Respostas sobre como os pesquisados avaliam o nível de conhecimento das ferramentas escolhidas.

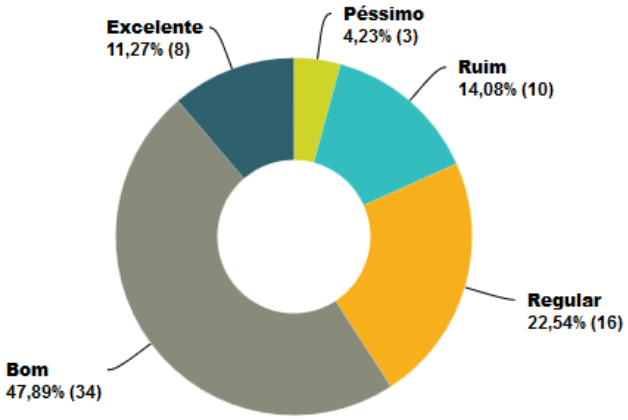


Figura 6. Respostas sobre como os pesquisados avaliam a utilidade e aplicabilidade das ferramentas à atividade de investigação

3.3.4.2 Análise dos Resultados

Os dados coletados, organizados e analisados serão apresentados em gráficos a seguir, juntamente com as observações.

Sobre a questão de pesquisa 1 (RQ1), quando os entrevistados foram questionados em relação às ferramentas utilizadas para Relatórios e Análises de Dados (OLAP/BI), dentro do grupo de respondentes que utilizam alguma ferramenta para Data Mining e Data Analytics ou BI, 44% dos respondentes utilizam o Excel, seguido do Microstrategy (16,82%), SAS (10,28%), Qlik-Qlikview (8,41%) e Oracle BI (3,74%) (vide Figura 7).

De igual modo, ainda em relação à questão de pesquisa supracitada, do rol de ferramentas propostas para os pesquisados, há uma série de ferramentas de análise e mineração de dados que, muito embora os órgãos possuam em suas instalações, não são utilizadas em seus processos investigativos de alguma forma, como visto na Figura 8.

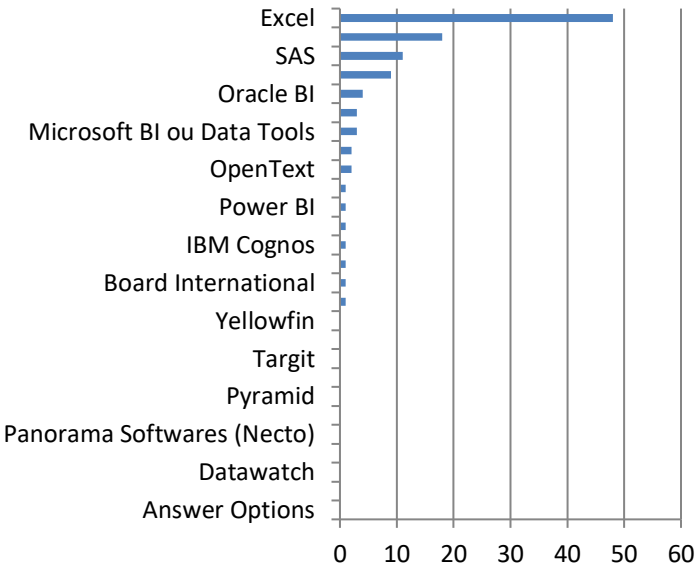


Figura 7. Respostas sobre ferramentas utilizadas para Relatórios e Análises de Dados (OLAP/BI).

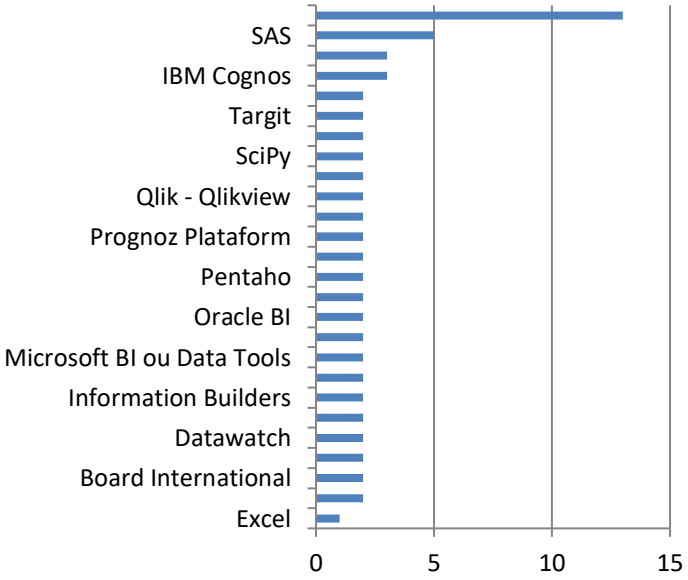


Figura 8. Respostas sobre ferramentas para Relatórios e Análises de Dados (OLAP/BI) que estão disponíveis no ambiente de trabalho, mas não são usadas.

Para a questão de pesquisa 2 (RQ2), foi questionado ainda sobre o período de experiência e utilização das ferramentas. Como pode ser visto na Figura 9, se

considerarmos os usuários com menos de três anos, dentre aqueles que afirmaram utilizar alguma ferramenta, há pouca experiência nos produtos, 64% dos investigadores.

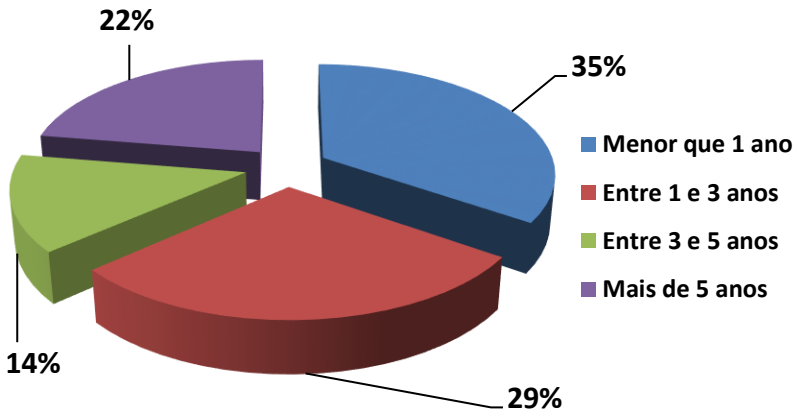


Figura 9. Respostas sobre experiência com as ferramentas utilizadas para Relatórios e Análises de Dados (OLAP/BI).

Para responder à questão de pesquisa 3 (RQ3), foi utilizada uma pergunta para identificar como os usuários utilizam os recursos das ferramentas. A Figura 10 expõe os dados apontados sobre a escolha dos recursos utilizados, a partir das ferramentas disponibilizadas, ou seja, aquelas que são utilizadas na geração de relatórios investigativos. De fato, 57,50% dos respondentes apontam desconhecer esse tipo de tecnologia.

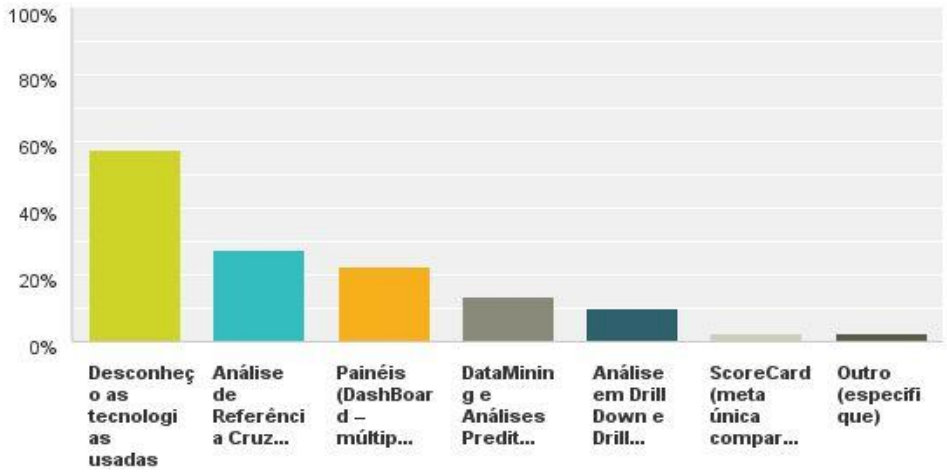


Figura 10. Respostas sobre recursos utilizados para Relatórios e Análises de Dados (OLAP/BI).

O recurso mais utilizado é a Análise de Referência Cruzada, representando 27,50% das respostas, seguido da opção Painéis (DashBoard – múltiplas métricas presentes visualmente), a qual representa 22,50%.

Para responder à questão 4 (RQ4) da pesquisa em apreço, como também a hipótese proposta e descrita no planejamento inicial, foi questionado aos pesquisados quais as técnicas ou algoritmos de inteligência artificial e Data Mining são utilizadas em seu trabalho. As respostas podem ser observadas na Tabela 1.

Para testar a hipótese em tela, utilizando o SPSS [09], software da IBM para análise estatística, passamos a testar a independência das variáveis frequência de agências e frequência de não utilização de algum algoritmo inteligente. Para tanto, aplicamos o teste de correlação de Pearson [10], o qual pode ser utilizado em variáveis qualitativas, quando se deseja comparar as distribuições de frequências obtidas contra as frequências esperadas. A hipótese nula testada é: "as variáveis são independentes" (vide H0, no planejamento). Portanto, no caso de encontrada diferença significativa, deve-se rejeitar a hipótese nula em favor da hipótese alternativa: "as variáveis não são independentes" (vide H1, no planejamento), conforme nível de significância acostado na tabela (vide Figura 11).

Tabela 1. Respostas sobre quais as técnicas ou algoritmos de inteligência artificial e *Data Mining* são utilizadas no trabalho.

Técnicas	**Porcentagem do total (108)**
Detecção de anomalias (*outliers*)	2,77% (3)
Agrupamento (clusterização)	1,85% (2)
Análise de componentes principais	1,85% (2)
Modelos de regressão	1,85% (2)
Regras de associação	0,92% (1)
Modelos probabilísticos (Naives Bayes, etc.)	0,92% (1)
Árvores de decisão	0,92% (1)
Reconhecimento de face ou Imagens	0,92% (1)
k-Nearest Neighbours (KNN)	0,00% (0)
Deep Learning	0,00% (0)
Reconhecimento de fala	0,00% (0)
Redes Neurais	0,00% (0)
Máquina de vetores de suporte	0,00% (0)

Para este objeto de estudo, a rejeição da hipótese nula (H0), a um nível de significância de 0,01, representará a existência de evidências estatísticas de que quanto maior o número de unidades especializadas em investigações complexas, menor é o uso de técnicas de Data Mining, ou, em um raciocínio inverso, maior é a não utilização destas técnicas. As frequências das respostas foram utilizadas como entrada para o teste. O resultado está apontado na saída do SPSS, constante na Figura 11.

A correlação de Pearson revelou um p-value de 0,0001, abaixo do nível de significância adotado, concluindo-se que devemos rejeitar a hipótese nula. Desta forma, há uma associação extremamente forte entre as variáveis e a evidência do pouco uso de inteligência computacional nas investigações, corroborando ou derrocando pressupostos investigativos. Tal fato pode estar relacionado à falta de conhecimento específico sobre os conceitos e o uso das referidas técnicas.

Correlações

		VAR00001	VAR00002
VAR00001	Correlação de Pearson	1	1,000**
	Sig. (2 extremidades)		,000
	N	108	108
VAR00002	Correlação de Pearson	1,000**	1
	Sig. (2 extremidades)	,000	
	N	108	108

**. A correlação é significativa no nível 0,01 (2 extremidades).

Figura 11. Correlação de Pearson.

Na mesma linha, é possível abstrair uma segunda justificativa: as aquisições de softwares e aplicações em tecnologias investigativas, feitas por órgãos públicos, são, em sua grande maioria, acompanhadas de treinamento nas referidas aplicações, no entanto, a experiência dos autores deste artigo tem mostrado que os modelos utilizados nos exemplos e nos exercícios, durante os treinamentos, têm origem no mercado privado, sem emular situações reais e realizar estudos de caso com os dados oriundos das investigações mais complexas. Este fato pode sugerir uma das causas para esse resultado, a qual poderá ser aprofundada em outros artigos.

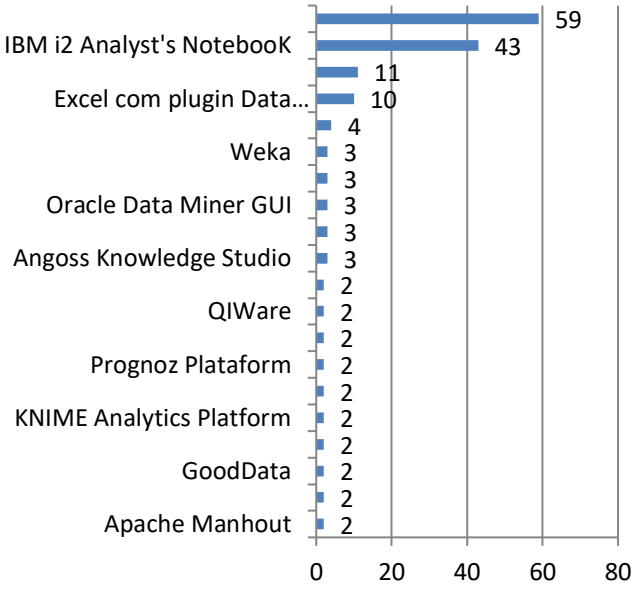


Figura 12. Respostas sobre quais ferramentas de mineração de dados (Data Mining) estão disponíveis no seu local de trabalho para auxiliar nas investigações e há quanto tempo vêm sendo usadas.

Ainda nesse contexto, na Figura 12, com relação ao uso de ferramenta específica de mineração de dados, 57% não souberam responder a questão, 41% utilizam o IBM i2 *Analyst's Notbook*, 10,68% utilizam o IBM Watson e 9,70% utilizam o Excel com Plugin Data Mining. Dentro desta mesma questão, 37,86% dos respondentes, muito embora possuam essas ferramentas em suas unidades investigativas, não as utilizam, o que, de fato, é um percentual bastante elevado.

Entre as técnicas respondidas e que estão relacionadas na Tabela 1, a detecção de outliers é mais usada, perfazendo um número de apenas 3 respondentes (2,77% do total). Estes dados dão um indício inicial de que as técnicas são muito pouco conhecidas por aqueles que executam funções investigativas com manipulação de dados e uso de recursos computacionais.

3.4 Ameaças à Validade

Alguns problemas podem ser ocasionados durante a participação dos indivíduos no questionário:

- Instrumentação adequadamente preparada para a execução (validade interna): Os participantes responderam ao questionário sem nenhuma

supervisão, assim, há a probabilidade dos mesmos não terem entendido uma questão específica. Para mitigar esse tipo de problema, um piloto foi realizado com 7 (sete) respondentes iniciais, de maneira a contribuir com modificações e focar na clareza das questões.

- Representatividade da população estudada (validade externa): A dificuldade em atingir os órgãos de inteligência estaduais, devido às suas naturezas investigativas e sigilosas, foi um grande desafio. Contudo, a grande necessidade que as agências estaduais possuem em radiografar o estado da arte das principais técnicas e tecnologias de investigação, em nível nacional, contribuiu para que todas as agências estaduais concordassem em colaborar e responder ao questionário proposto.
- Distribuição do conjunto de participantes (validade de conclusão): A expertise dos profissionais ou as suas funções podem afetar os resultados, contudo, a participação de todos os estados e a variabilidade das funções encontradas mitigaram esta ameaça e apoiaram a correlação encontrada, a qual foi testada estatisticamente.

3.5 Conclusões

O presente trabalho é a resultante de uma pesquisa quantitativa que pode ser utilizada por agentes de segurança pública, analistas do ministério público, coordenadores, promotores de justiça, delegados de polícia e gestores da segurança pública de forma geral, para a tomada de decisão, bem como por pesquisadores da área, para direcionar suas pesquisas nesta lacuna investigativa aqui apresentada. Diante dos dados coletados, ficou constatada a capilaridade dos respondentes, abrangendo as polícias de forma geral e os ministérios públicos estaduais e federal.

Com relação aos bancos de dados utilizados, 48% das agências usam Oracle e 42,50% usam SQL Server, o que indica um domínio do uso de tecnologias proprietárias ou ainda uma falta de domínio em tecnologias de código aberto. No que diz respeito ao processo de integração via ETL, 40% não conheciam o processo, um percentual muito expressivo, diante de um procedimento tão importante e significativo para efetividade e

qualidade nas cargas dos dados. Outro ponto de destaque é o fato de 97% dos respondentes não utilizarem técnicas de mineração de dados. Além disso, 30,99% avaliam o próprio conhecimento sobre as ferramentas que utilizam como ruim ou péssimo.

Esses dados expõem uma realidade dura dentro dos principais órgãos de investigação e controle do nosso país, pois apesar de identificarmos o uso de várias ferramentas, ainda falta a exploração adequada, o que pode sugerir um investimento muito grande por parte do estado na compra e aquisição destes artefatos, não obstante, sem o retorno desejado. Também fica evidente a falta de capacitação adequada e de conhecimento suficiente por parte dos envolvidos no uso dos softwares, bem como no uso das técnicas que servem como principais características dessas aplicações.

A principal dificuldade deste trabalho foi a difícil tarefa da aplicação de uma pesquisa do tipo "Survey" dentro dos órgãos de inteligência e dos laboratórios de tecnologias de combate à lavagem de dinheiro, por conta do sigilo investigativo que serve de referência e de modelo para o país, bem como pela própria natureza de trabalho destes órgãos. Neste contexto de seriedade e sigilo, houve a preocupação constante com a veracidade nas respostas dadas pelos participantes, a qual foi contornada com o acesso pessoal ou por telefone a cada um dos respondentes, mostrando as necessidade e importância do estudo. Todo este esforço proporcionou o atingimento de todo o território nacional.

Como trabalhos futuros, sugere-se aprofundar a pesquisa e coleta de informações junto aos órgãos envolvidos, com o fito de entender quais as causas dos parcos uso e entendimento dos softwares já disponíveis, bem como compreender o pouco uso de tecnologias e softwares livres. Destaca-se, também, a necessidade de ampliar as pesquisas com o foco em encontrar lacunas ainda não identificadas neste Survey.

Por fim, destacamos a relevância desta pesquisa e dos órgãos aqui envolvidos, os quais têm a responsabilidade social de apresentar resultados que servem como direcionamento para as decisões judiciais em todo o país. A proeminência destes órgãos e este Survey devem servir de alerta e direcionamento para os nossos governantes.

3.6 Referências

[1] Brasil. Lei nº 9.883, de 07 de Dezembro de 1999. Institui o Sistema Brasileiro de Inteligência, cria a Agência Brasileira de Inteligência – ABIN, e dá outras providências.

[2] Brasil. Ministério da Justiça. Secretaria Nacional de Segurança Pública. Doutrina Nacional de Inteligência de Segurança Pública. Brasília, 2014.

[3] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau. Crime Data Mining: a general framework and some examples. *Computer*, 37(4):50- 56, 2004.

[4] J. K. G. Costa, I. P. O. Santos, M. C. Junior, and A. V. R. Nascimento. Um experimento em um ambiente de business intelligence industrial para melhoria da manutenção de cargas de dados. *SBSI*, 2016.

[5] I. H. Witten, E. Frank, M. A. Hall, and C. J. *Pal. Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2016.

[6] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton. Big data. *The management revolution*. Harvard Bus Rev, 90(10):61-67, 2012.

[7] V. R. Basili and D. M. Weiss. *A methodology for collecting valid software engineering data. Technical report*, DTIC Document, 1983.

[8] R. Van Solingen and E. Berghout. *The Goal/Question/Metric Method,* McGraw-Hill, 1999.

[9] SPSS Inc. Released 2017. SPSS for Windows, Version 24.0. Chicago, SPSS Inc.

[10] R. L. Plackett. Karl pearson and the chi-squared test. *International Statistical Review/Revue Internationale de Statistique*, pages 59-72, 1983.

[11] R. Kimball and M. Ross. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

[12] F. J. Braz, W. S. Coan, and A. Rosseti. Uma proposta de solução de mineração de dados aplicada à segurança pública. SBSI, 2012.

[13] V. Lourenço, P. Mann, A. Paes, and D. de Oliveira. Siapp: Um sistema para análise de ocorrências de crimes baseado em aprendizado lógico-relacional. SBSI, 2016.

[14] A. B. Leite, E. P. R. Souza, J. d. S. C. Neto, and M. I. de Sousa Oliveira. Aplicação olap para segurança pública: um estudo de caso a partir de dados governamentais abertos do estado do Rio de Janeiro. São Paulo. SBSI, 2012.

CAPÍTULO 4

4 ARTIGO II - MAPEAMENTO

Fighting Against Money Laundering: A Systematic Mapping

Bruno Luiz Kreutz Barroso *, Fabio Mangueira´† and Methanias Colaço Junior´‡
Postgraduate Program in Computer Science - PROCC
Federal University of Sergipe (UFS)
São Cristóvão ao, Sergipe, Brazil˜
*kreutz.dev@gmail.com, †fabio.dipolcgi@gmail.com, ‡mrjse@hotmail.com

***Abstract*— Context:** Money Laundering (ML) is a global crime that has a close relation with other crimes, such as: illegal drug trading, terrorism or arms trafficking. Criminals in today's technology-driven society use every means available at their disposal to launder the profit made from their illegal activities. In response, international anti-money laundering (AML) efforts are made with AML systems. **Objective:** Identify and systematize the approaches, techniques and algorithms used in Computer Science (CS) to fight ML, besides identifying the trends in the field. **Method:** A systematic literature mapping was conducted to analyze the scientific research in the field. Results: The main approaches were identified, supervised classifiers and clusters, along with the trend of papers published over the years. China was the country with the highest number of published papers. **Conclusion:** The most relevant studies in such research line adopt data mining and machine learning techniques using clusters and classifiers. The state of the art was mapped, making it clear that it is an area of interest for researchers around the world with growth potential. We believe that this work is relevant to the academy, governments and the community at large, presenting them with trends in the detection of money laundering.

Index Terms—Money Laundering, Data Mining, Artificial Intelligence, Machine Learning, Cyber Crimes and Profiling.

4.1 INTRODUCTION

Money laundering (ML) usually refers to such activity or processes that deals with criminal proceeds to disguise their illicit origin and make them look legit [1]. ML is considered as a major crime in criminology, and is identified as one of the top group crimes in today's society [2], besides being, frequently, a transnational crime that occurs

in close relation to other crimes, like illegal drug trading, terrorism, or arms trafficking [3].

Criminal elements in today's technology-driven society use every means available at their disposal to launder the proceeds from their illegal activities. In response, international community has made anti-money laundering (AML) efforts are being made [4]. Usually, financial institutions use semiautomated processes to flag suspicious ML transactions, based on medians and predetermined standard irregularities [5]. AML systems are pivotal and fundamentals to aid governments and institutions to fight against ML.

In this context, is necessary to identify the best practices to combat ML, the best techniques and opportunities to be explored. It is needed to disseminate a culture of repression to this kind of delict, which accompany, encourages and finances the apparatus and investment of many other daily delicts, presenting a dangerous threat to the society.

This article presents a Systematic Mapping that had as objective to identify and systematize the approaches, techniques and algorithms used to detect ML. With this purpose, articles from important databases of CS were mapped.

After answering the research questions it was identified that the main techniques explored were supervised classification techniques [1], [5]–[18] with 15 (28.3%) and clustering [2], [5], [7], [10], [15], [17], [19]–[25] with 14 techniques (26.42%).

As for the characterization of the publications, the amount oscillated a lot over the years, mostly because of the low amounts of publications. In relation to the countries, China was, by a large margin, the country with the most publications in the field. The peak of publications about the theme was in 2010, the IEEE International Conference on Machine Learning and Applications (ICMLA), the International Conference on Machine Learning and Cybernetics (ICMLC) and the IEEE International Conference on Data Mining Workshops (ICDMW) published the most papers. The conferences dominated the publications landscape. Finally, two similar papers from the same city, published by two different authors were identified.

This paper is organized as follows: in section 2, the literature works related to the theme of this systematic mapping are presented; in section 3, the method adopted in this mapping is presented; in section 4, the results of the analysis are described; in section 5, threats to validity are presented; Finally, in section 6, the conclusion is presented.

4.2. RELATED WORKS

Secondary studies related to our research were found. Ngai et al. [26] presented a classification framework and a systematic review on the application of data mining techniques in the detection of financial fraud. D. Yue et al. [27], also presented a generic framework for understanding and classifying different combinations of financial fraud detection techniques and data mining algorithms. However, unlike the present study, the work referred [26] and [27] were not just about money laundering, this type of crime was only a subset of the financial fraud classified. In the systematic review conducted by Ngai et al. [26], only one primary study dealing with money laundering was found.

This paper distinguishes itself by treating and focusing itself solemnly in money laundering, by not focusing solely on data mining and by emphasizing the techniques of outlier detection and time series. In addition, the present mapping contemplates more recent studies.

4.3 METHOD

Some researchers have been working to establish stable methods for applying the systematic review process in the literature [28]–[31]. One of these methods is the Systematic Mapping, which consists of a systematic protocol for searching and selecting relevant studies in the literature, with the objective of extracting information and mapping the results to a specific research problem [28], [29]. The present study was based on the protocols proposed by Kitchenham et al. [28] and Petersen et al. [29].

The choice of performing Systematic Mapping was justified by allowing the analysis of primary studies in a broader way to answer the research questions, as well as collecting evidence to guide future research.

A. *Research Questions*

The objective of this paper was to perform a Systematic Mapping with the purpose of identifying and analyzing primary studies, to characterize the use of algorithms, methods and techniques to detect evidence of money laundering. For the research questions' elaboration, initially, it was decided to detail the approaches used for the detection of outliers and time series. This approach was based on control papers and the assumption that the problem of money laundering produces data that favors the discovery of anomalies, since the detection of suspicious activities can be seen as a outlier detection problem [22]. In addition, time series are used in the scope of several financial problems [32]. In this context, the study intended to highlight the researches that used these two techniques, while not failing to identify all the others. Furthermore, it was intended to identify the current panorama of the ML detection field, in order to guide future research. Thus, the following questions were elaborated:

• Q1: What are the most commonly used computational approaches and techniques as basis for money laundering detection?

• Q2: What specific outlier discovery methods or algorithms are used to identify transactions that may indicate money laundering?

• Q3: What specific time series analysis methods or algorithms are used to identify transactions that may indicate money laundering?

• Q4: Which countries have the highest number of research published on this context?

• Q5: Which years have had the most publications in this area?

• Q6: What are the main journals and conferences about the subject?

• Q7: Which is the most popular publication venue?

B. *Search Strategy*

The following bases were used to execute the Systematic Mapping: Scopus, IEEE and ACM. Download without restriction was granted through the Capes journals portal (https://www.periodicos.capes.gov.br). The Scopus base was chose due its comprehensively collection of articles from several databases: Science Direct, Springer and Elsevier are among them [33]. To supplement the Scopus results the ACM and IEEE

bases were used. These databases are responsible for publishing the major journals and conferences in the area of CS.

Sources were selected through the keywords search according to their availability on the internet. Only English studies, works related to CS and articles published in conferences, periodicals or book chapters were selected.

The advanced search refinement option was used in the

Scopus database to select only results within the field of CS whose language were English. Also, results referring to conference recapitulations and notes were excluded. In the other bases no refinement was made.

The search string used, generated with the keywords, was: ((”money laundering” OR ”capital laundering”) AND (”data mining” OR ”data analytics” OR ”outlier” OR ”forecasting” OR ”time series” OR ”big data” OR ”business intelligence” OR ”data science” OR ”artificial intelligence” OR ”machine learning”))

With the search conducted during August – November of 2017, 86 unique results were returned by the search strings. After this stage, the papers selection was started, which will be detailed below.

C. *Selection Criteria*

In order to filter the relevant papers to this Systematic Mapping, the inclusion and exclusion criteria were established. The study used the following inclusion criteria:

• The result should contain the theme of this study in the title, abstract or keywords;

• The result needs to explore an algorithm, technique, mechanism or approach for money laundering detection.

To confirm the inclusion criteria, the abstract and introduction of each paper were analyzed.

In parallel, the articles were analyzed according to the exclusion criteria. The exclusion criteria described below was also applied to them:

• Papers that do not belong to the field of CS;

• Secondary studies, as they deal with third-party approaches;

• Papers that were unavailable;

• Ongoing studies.

After the inclusion and exclusion criteria were applied, the relevance of the studies were evaluated. Among the 86 unique papers found, 35 were selected to compose the primary studies. As 2 of these 35 articles ([34] and [35]) were considerably similar but had distinct authors, it was decided to count the two articles as one to avoid noise caused by the duplication of one of the studies, totaling, in the end, 34 primary studies. Therefore, when one of these articles is referred in this paper it means both are being referenced. Nevertheless, it is not the aim of this study to prove that there was any unethical behaviour or violation, further investigation, perhaps by IEEE and Springer, would be necessary, as they could contact the authors and give them the right to defend themselves if they judge there was any misconduct.

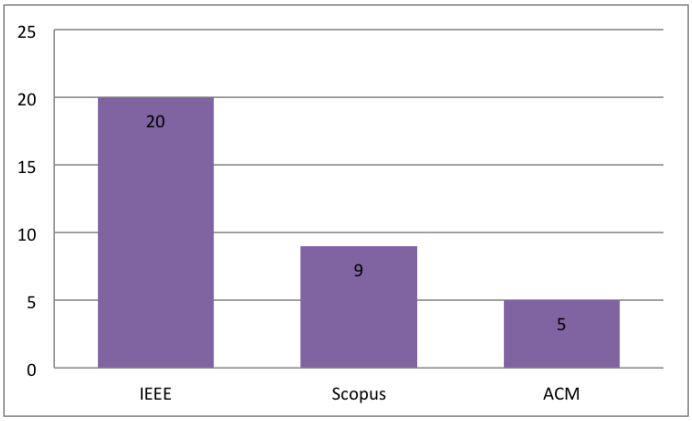


Figure 1. Articles selected by base

The chart in the figure 1 shows the amount of articles by scientific repository after the application of the selection criteria. Although, IEEE, ACM are under the Scopus umbrella, the papers counted as Scopus were the ones found in other bases.

4.4. DISCUSSION

In this section, we present the analysis results of the primary studies, answering the research questions presented earlier.

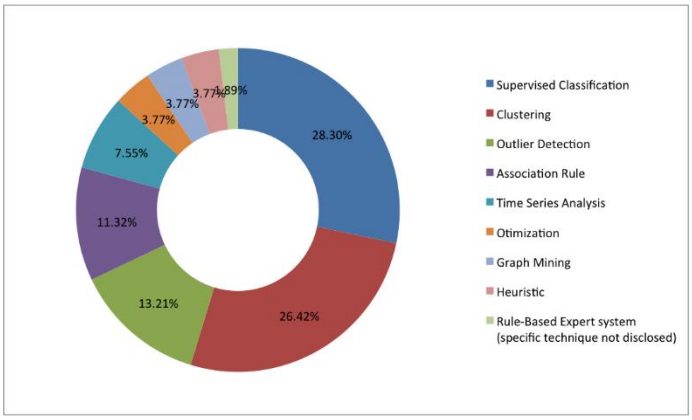


Figure 2. Characterization

In the chart displayed in figure 2, is presented the characterization of the main approaches found for Q1. The most relevant the techniques were supervised classification techniques [1], [5]–[18] with 15 instances (28.3%), followed by the ones based on clustering [2], [5], [7], [10], [15], [17], [19]–[25] with 14 (26.42%) of the main techniques from the papers. Outlier detection techniques [21], [22], [34], [36], [37] had 7 (13.21%), while association rules techniques [5], [24], [38] had 6 algorithms (11.32%). Time series analysis [2], [32], [39], [40] is the next with 4 (7.55%). Graph mining [41], [42], optimization algorithms [1], [36] and heuristics [7], [23] had 2 (3.77%) each. Finally, the paper that used a rule-based expert system did not disclose the specific algorithm used [10], representing 1 instance (1.89%).

Among the primary studies, 5 proposed the use of intelligent agents [4], [5], [13], [43], [44], but only 2 of them specified the techniques implemented: clusters, supervised classifiers and association rules [5], and supervised classifiers [13]. The others didn't go into the algorithm details.

Supervised classification builds up and utilizes a model to predict the categorical labels of unknown objects to distinguish between objects of different classes [26], [45]. Amid the supervised classifiers, decision trees had the most occurrences (7 out of 15, 46.66%), followed by neural networks (4 out of 15, 26.66%) and SVM (3 out of 15, 20%). The drawback of supervised classifiers is that they need labeled data to be trained and the process of labeling data when dealing with big data sets may be exhaustive.

Clustering is used to divide objects into conceptually meaningful groups (clusters), with the objects in a group being similar to one another but very dissimilar to the objects in other groups. Clustering is also known as data segmentation or partitioning and is regarded as a variant of unsupervised classification [26], [45], [46]. Clusters prominent use is due to their ability find meaningful structures in the data set. Some studies didn't specify the clustering algorithm implemented, using terms as: "centre-based clustering algorithm", "proprietary clustering algorithm " and "modified algorithm", but in the ones that did, K-Means was the most popular with 3 instances in 14 (21.42%), followed by Improved Minimum Spanning Tree clustering Algorithm, DBSCAN, CLOPE, EM and CBLOF with 1 instance each (7.42%). The other clustering techniques were not specified.

The answer to Q2 is presented in Table I. Anomaly detection can identify unexpected activity in the regular data-flow [47]. Outlier detection is employed to measure the "distance" between data objects to detect those objects that are grossly different from or inconsistent with the remaining data set [26], [45]. In this table, the outlier detection techniques were elucidated. The 7 outlier detection algorithms identified are: Cross Dataset Outlier Detection Model, Dissimilarity Metric, Isolation Forest, One class SVM, Gaussian Mixture Model, Hidden Markov Model and Local Outlier Factor. All algorithms had only one instance in the papers.

The response to Q3 is presented in table II. Time series analysis comprises methods for analyzing a sequence of data points, measured typically at successive time spaced uniform intervals, in order to extract meaningful statistics and other characteristics of the data [48]. Time series analysis algorithms that were presented are further detailed in this table.

Table I ALGORITHMS AND TECHNIQUES

Outilier Detection Technique	Reference
Dissimilarity Metric	[21]
Cross Dataset Outlier Detection Model	[34]
Isolation Forest	[37]
One class SVM	[37]
Gaussian Mixture Model	[37]

Hidden Markov Model	[36]
Local Outlier Factor	[22]

All algorithms, methods or techniques for time series analysis were found only once, being: Time variant behavioral pattern, Sequence Matching Based Algorithm, Scan Statistics Based Method and Correlation Analysis Along Timeline.

Table II
ALGORITHMS AND TECHNIQUES (TIME SERIES)

Time Series Technique	Reference
Time variant behavioral pattern	[39]
Sequence Matching Based Algorithm	[32]
Scan Statistics Based Method	[40]
Correlation Analysis Along Timeline	[2]

The answer to Q4 is shown in figure 3. The amount of papers per country was counted using the affiliation of the authors as parameter, if the authors were affiliated to institutes from different countries both countries were counted. China leads the number of publications, with a total of 17 papers. Australia, Poland, United States, India and Ireland appear next, with 2 publications each. The other countries, Brazil, Hong Kong, Portugal and Vietnam, United Kingdom, Malaysia, Canada and Luxembourg have 1 publication each.

China's leadership may be due to government policy changes in relation to the financial system, which began in 1976, when it was virtually non-existent and improved in the following decades with the increased role of independent financial activity [49], perhaps Chinese government may be more eager to fund researches on this topic. Another more obvious hypothesis, in this context, is that China may simply have a larger number of researchers working on data mining and artificial intelligence applications than other countries.

The answer to Q5 is presented in figure 4, in which it can be observed that the year with the highest number of publications was 2010, with 6 papers. In 2009, 5 papers were published and in 2014, 4. The oldest publication on the subject dates back to 2003. There

have been an year in which no paper was published: 2013. It is notorious the scarcity of publications on the field and the oscillations over the years.

In the above chart, figure 4, the distribution of publications per venue of publication can also be observed. The only publication whose primary source was a book chapter, occurred in the year in which the number of publications from conferences was higher, 2010. The conferences accounted for the largest number of publications in almost every year, except in 2014, year in which the peak of publications in journals occurred and tied the number of publications from conferences in that year.

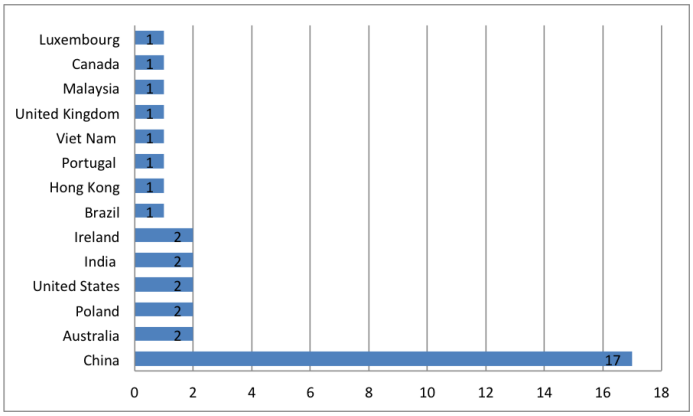


Figure 3. Papers per Country

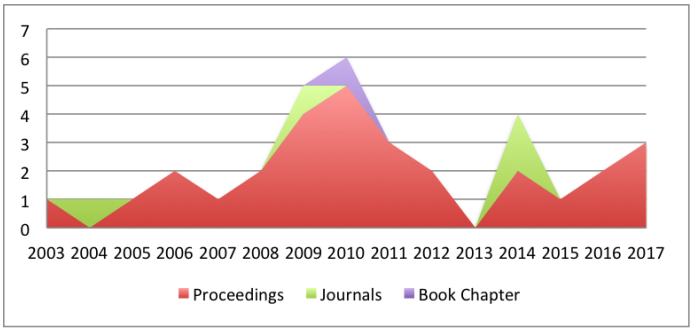


Figure 4. Papers per Year

The answer to Q6 is that the ICMLA, ICMLC, ICDMW were the main conference identified with each one being the source of 2 publications. All other publications from conferences were originated from different sources. The journals were the source of 1 paper each: Expert Systems with Applications, IEEE Intelligent Systems, International Journal of Security and Applications and Journal of Theoretical and Applied Information Technology.

The low absolute number of papers published in the main conferences identified indicates that others can catch up to them in the near future. It also indicates that there may be a lack of conferences dedicated exclusively to fraud detection techniques in general, probably because the highly specificity of the field.

Finally, the answer to Q7 is shown in figure 5, showing that the most popular venue to publish papers on this topic are conferences. This pattern is not surprising, since the most accessible medium for scientific publications are known as conferences. For example, over 100,000 conference events worldwide are indexed in the Scopus database, whilst only nearly 22,000 journals are indexed [33]. Surprising is the low absolute number of publications found in journals, which may denote that papers published at conferences have not been sufficiently worthy for their extension in journals, that is, they may not have been of sufficient quality. In addition, one possibility is that the results were not instigating enough for the deepening and continuity of the studies.

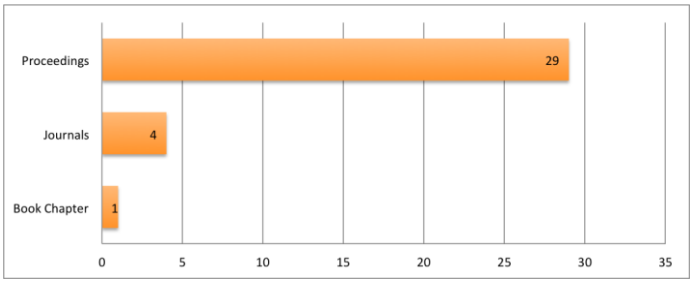


Figure 5. Publication Distribution

4.5 THREATS TO VALIDITY

Construction Validity: The search string may not cover the whole money laundering detection area. To mitigate this threat, we sought to construct the most comprehensive string possible utilizing a control paper and the opinion of one of the three researchers, a member of government staff that investigates Money Laundering.

Internal validity: (Data extraction): Researchers were responsible for extracting and classifying the main algorithms of each publication, biases or data extraction problems can threaten the validity of data characterization; (Selection Bias): Some papers may have been categorized incorrectly as the articles were included or excluded in the systematic mapping according to the researchers' judgment.

To mitigate these threats, selection and extraction reviews were made by all three researchers involved, with a final vote on disagreements.

External Validity: Although Scopus is the largest database of scientific literature, with over 60 million records and 21,500 journals [33], it's impossible to state that the results of this systematic mapping covered all of CS. Nevertheless, this study presented evidence of the main techniques used and gaps to be explored, serving as a guide for future works in this line.

4.6 CONCLUSION

In this work, a systematic mapping was carried out, aiming to identify scientific papers related to the analysis and evaluation of algorithms, methods and techniques in the field of CS, to detect and combat ML. Since no other similar work of Mapping or Systematic Review specifically about ML has been found, we assumed that this is the first work of this type in this specific area of academic scientific knowledge.

This mapping was conducted following the research protocol and selection of studies presented in section 2. With this method, data from 34 primary studies was extracted and analyzed, identifying trends in this area.

As results, it was identified that the most relevant techniques identified were supervised classifiers, firstly, and clusters, secondly. Between the former, decision trees (first), neural network and SVM were the most used algorithms and, amid the clusters, K-Means, first of all, followed by Improved Minimum Spanning Tree clustering Algorithm, DBSCAN, CLOPE, EM and CBLOF were the main ones ((Q1).

There was no repetition between the algorithms used for Outlier detection and Time series analysis, which indicates that there is no consolidate approach for both. Thus, all algorithms classified as Outlier detection and Time series analysis in the studies were highlighted (Q2, Q3).

In the global scenario, China stands out firstly, followed by Poland, Australia, Ireland, India and the United States, all in second place, as the countries that have published the most papers, presenting, respectively, 17, 2, 2, 2, 2 and 2 publications each (Q4).

Over the years, the number of publications fluctuates a lot, the oldest publication dates back to 2003. The year with the most publications was 2010, when 4 papers were published. It is notorious the scarcity of publications in this field of research (Q5).

The ICMLA, ICMLC and ICDMW were the main conferences identified, each one publishing 2 of the primary studies, while the main journals had one publication each. In this case, the low absolute number of papers published indicates that other journals and conferences can challenge their spot in the near future (Q6).

Finally, the most popular venue for publications on the topic are conferences, as expected, as the number of conferences indexed in the databases used is greater than the number of scientific journals (Q7).

Besides the apparent need to deepen the researches in the discussed area, the results found in this work map the state of the art of detecting transactions suspicious of money laundering, making it clear that it is an area of interest for researchers around the world and it has great growth potential. We believe that this work is relevant to the academy, governments and the community at large, presenting them with trends in the detection of money laundering. In addition, it can offer yet another approach in the search for the best solutions to the current scenario and to combat the threat of organized crime against society.

4.7 REFERENCES

[1] L.-T. Lv, N. Ji, and J.-L. Zhang, “A rbf neural network model for antimoney laundering,” in Wavelet Analysis and Pattern Recognition, 2008. ICWAPR’08. International Conference on, vol. 1. IEEE, 2008, pp. 209–215.

[2] Z. M. Zhang, J. J. Salerno, and P. S. Yu, “Applying data mining in investigating money laundering crimes,” in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003, pp. 747–752.

[3] P. A. Schott, Reference guide to anti-money laundering and combating the financing of terrorism. World Bank Publications, 2006.

[4] S. Gao and D. Xu, “Conceptual modeling and development of an intelligent agent-assisted decision support system for anti-money laundering,” Expert Systems with Applications, vol. 36, no. 2, pp. 1493–1504, 2009.

[5] C. Alexandre and J. Balsa, “Integrating client profiling in an anti-money laundering multi-agent based system,” in New Advances in Information Systems and Technologies. Springer, 2016, pp. 931–941.

[6] X. Luo, “Suspicious transaction detection for anti-money laundering,” Int. J. Secur. Its Appl, vol. 8, 2014.

[7] N. A. Le Khac, S. Markos, and M.-T. Kechadi, “A data mining-based solution for detecting suspicious money laundering cases in an investment bank,” in Advances in Databases Knowledge and Data Applications (DBKDA), 2010 Second International Conference on. IEEE, 2010, pp. 235–240.

[8] C. Ju and L. Zheng, “Research on suspicious financial transactions recognition based on privacy-preserving of classification algorithm,” in Education Technology and Computer Science, 2009. ETCS’09. First International Workshop on, vol. 2. IEEE, 2009, pp. 525–528.

[9] I. George and M. Kavakli, “Data mining in the investigation of money laundering and terrorist financing,” Surveillance Technologies and Early Warning Systems: Data Mining Applications for Risk Detection: Data Mining Applications for Risk Detection, p. 228, 2010.

[10] R. S. Freedman and I. Sobkowski, “Surveillance of parimutuel wagering integrity using expert systems and machine learning.” in IAAI, 2010.

[11] J. Schmidhuber, “Deep learning in neural networks: An overview,” Neural networks, vol. 61, pp. 85–117, 2015.

[12] S.-N. Wang and J.-G. Yang, “A money laundering risk evaluation method based on decision tree,” in Machine Learning and Cybernetics, 2007 International Conference on, vol. 1. IEEE, 2007, pp. 283–286.

[13] J. Kingdon, “Ai fights money laundering,” IEEE Intelligent Systems, vol. 19, no. 3, pp. 87–89, 2004.

[14] L. Keyan and Y. Tingting, “An improved support-vector network model for anti-money laundering,” in Management of e-Commerce and eGovernment (ICMeCG), 2011 Fifth International Conference on. IEEE, 2011, pp. 193–196.

[15] N. A. Le Khac and M.-T. Kechadi, “Application of data mining for antimoney laundering detection: A case study,” in Data Mining Workshops (ICDMW), 2010 IEEE International Conference on. IEEE, 2010, pp. 577–584.

[16] J. Tang and J. Yin, “Developing an intelligent data discriminating system of anti-money laundering based on svm,” in Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on, vol. 6. IEEE, 2005, pp. 3453–3457.

[17] R. Liu, X.-l. Qian, S. Mao, and S.-z. Zhu, “Research on anti-money laundering based on core decision tree algorithm,” in Control and Decision Conference (CCDC), 2011 Chinese. IEEE, 2011, pp. 4322– 4325.

[18] E. L. Paula, M. Ladeira, R. N. Carvalho, and T. Marzagao, “Deep˜ learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering,” in Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on. IEEE, 2016, pp. 954–960.

[19] D. K. Cao and P. Do, “Applying data mining in money laundering detection for the vietnamese banking industry,” in Asian Conference on Intelligent Information and Database Systems. Springer, 2012, pp. 207–216.

[20] Y. Yang, B. Lian, L. Li, C. Chen, and P. Li, “Dbscan clustering algorithm applied to identify suspicious financial transactions,” in CyberEnabled Distributed Computing and Knowledge Discovery (CyberC), 2014 International Conference on. IEEE, 2014, pp. 60–65.

[21] X. Wang and G. Dong, “Research on money laundering detection based on improved minimum spanning tree clustering and its application,” in Knowledge Acquisition and Modeling, 2009. KAM’09. Second International Symposium on, vol. 2. IEEE, 2009, pp. 62–64.

[22] Z. Gao, “Application of cluster-based local outlier factor algorithm in anti-money laundering,” in Management and Service Science, 2009. MASS’09. International Conference on. IEEE, 2009, pp. 1–4.

[23] T.-M. Cheong and Y.-W. Si, “Event-based approach to money laundering data analysis and visualization,” in Proceedings of the 3rd International Symposium on Visual Information Communication. ACM, 2010, p. 21.

[24] P. Umadevi and E. Divya, “Money laundering detection using tfa system,” 2012.

[25] Z. Chen, A. Nazir, E. N. Teoh, E. K. Karupiah et al., “Exploration of the effectiveness of expectation maximization algorithm for suspicious transaction detection in anti-money laundering,” in Open Systems (ICOS), 2014 IEEE Conference on. IEEE, 2014, pp. 145–149.

[26] E. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun, “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature,” Decision Support Systems, vol. 50, no. 3, pp. 559–569, 2011.

[27] D. Yue, X. Wu, Y. Wang, Y. Li, and C.-H. Chu, “A review of data mining-based financial fraud detection research,” in Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on. Ieee, 2007, pp. 5519–5522.

[28] B. Kitchenham, “Procedures for performing systematic reviews,” Keele, UK, Keele University, vol. 33, no. 2004, pp. 1–26, 2004.

[29] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, “Systematic mapping studies in software engineering.” in EASE, vol. 8, 2008, pp. 68–77.

[30] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, “Lessons from applying the systematic literature review process within the software engineering domain,” Journal of systems and software, vol. 80, no. 4, pp. 571–583, 2007.

[31] C. Wohlin, P. Runeson, P. A. d. M. S. Neto, E. Engstrom,¨ I. do Carmo Machado, and E. S. De Almeida, “On the reliability of mapping studies in software engineering,” Journal of Systems and Software, vol. 86, no. 10, pp. 2594–2610, 2013.

[32] X. Liu, P. Zhang, and D. Zeng, “Sequence matching for suspicious activity detection in anti-money laundering,” Intelligence and Security Informatics, pp. 50–61, 2008.

[33] B. Elsevier, “Scopus content coverage guide,” 2017.

[34] T. Zhu, “An outlier detection model based on cross datasets comparison for financial surveillance,” in Services Computing, 2006. APSCC’06. IEEE Asia-Pacific Conference on. IEEE, 2006, pp. 601–604.

[35] T. Jun, “A cross datasets referring outlier detection model applied to suspicious financial transaction discrimination,” in Intelligence and Security Informatics. Springer, 2006, pp. 58–65.

[36] Y. Li, D. Duan, G. Hu, and Z. Lu, “Discovering hidden group in financial transaction network using hidden markov model and genetic algorithm,” in Fuzzy Systems and Knowledge Discovery, 2009. FSKD’09. Sixth International Conference on, vol. 5. IEEE, 2009, pp. 253–258.

[37] R. D. Camino, R. State, L. Montero, and P. Valtchev, “Finding suspicious activities in financial transactions and distributed ledgers,” in 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2017, pp. 787–796.

[38] R. Drezewski, G. Dziuban, Ł. Hernik, and M. Paczek, “Comparison˙ of data mining techniques for money laundering detection system,” in Science in Information Technology (ICSITech), 2015 International Conference on. IEEE, 2015, pp. 5–10.

[39] G. K. MCA, M. PHIL, and M. PRABAKARAN, “Money laundering analysis based on time variant behavioral transaction patterns using data mining.” Journal of Theoretical & Applied Information Technology, vol. 67, no. 1, 2014.

[40] X. Liu and P. Zhang, “A scan statistics based suspicious transactions detection model for anti-money laundering (aml) in financial institutions,” in Multimedia Communications (Mediacom), 2010 International Conference on. IEEE, 2010, pp. 210–213.

[41] K. Michalak and J. Korczak, “Graph mining approach to suspicious transaction detection,” in Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on. IEEE, 2011, pp. 69–75.

[42] X. Li, X. Cao, X. Qiu, J. Zhao, and J. Zheng, “Intelligent anti-money laundering solution based upon novel community detection in massive transaction networks on spark,” in Advanced Cloud and Big Data (CBD), 2017 Fifth International Conference on. IEEE, 2017, pp. 176–181.

[43] S. Gao, D. Xu, H. Wang, and Y. Wang, “Intelligent anti-money laundering system,” in Service Operations and Logistics, and Informatics, 2006. SOLI’06. IEEE International Conference on. IEEE, 2006, pp. 851–856.

[44] C. Alexandre and J. Balsa, “A multiagent based approach to money laundering detection and prevention.” in ICAART (1), 2015, pp. 230– 235.

[45] P.-N. Tan, M. Steinbach, and V. Kumar, “Introduction to data mining. 1st,” 2005.

[46] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

[47] Y. B. Reddy, “Event-based anomalies in big data,” in Information Technology-New Generations. Springer, 2018, pp. 33–42.

[48] R. A. K.-l. Lin and H. S. S. K. Shim, “Fast similarity search in the presence of noise, scaling, and translation in time-series databases,” in Proceeding of the 21th International Conference on Very Large Data Bases. Citeseer, 1995, pp. 490–501.

[49] A. Keidel, “China’s financial sector: Contributions to growth and downside risks,” in China’s Emerging Financial Markets. Springer, 2009, pp. 111–125.

CAPÍTULO 5

5 EXPERIMENTO

Análise Comparativa do EM e do K-Means Aplicados ao Contexto de Detecção de Transações Suspeitas em Investigações de Lavagem de Dinheiro

Este capítulo apresenta-se como uma sequência lógica, fruto do encadeamento proposto. Conhecer a realidade do Brasil, detectar as principais publicações no mundo e, por fim, realizar um experimento em ambiente real, similar a um dos trabalhos encontrados.

5.1 Trabalhos Relacionados

Na busca realizada nas bases de dados, foram analisados 20 artigos do IEEE, 9 da *Scopus* e 5 artigos da ACM. Assim, fizeram parte do escopo deste experimento, o estudo e análise de 34 artigos científicos (vide capítulo 4).

Considerando os trabalhos relacionados diretamente com este experimento, observou-se que pesquisas foram conduzidas para aplicar técnicas de agrupamento na identificação de transações suspeitas de lavagem de dinheiro, nas quais o uso do K-means ou de suas variações é proeminente (Gao, 2009; Le Khac, 2010; Umadevi, 2012; Chen et al., 2014).

No trabalho de Chen et al. (2014), Expectation Maximization (EM) e K-Means foram usados para detectar transações suspeitas, agrupando os dados por espaços de tempo diário, semanal e mensal. Em suas conclusões, Chen et al. (2014) evidenciaram que o algoritmo EM teve um comportamento melhor que o K-Means, no que concerne à detecção de lavagem de dinheiro. Ainda segundo os autores, foi utilizada uma base de dados com um conjunto de informações bancárias reais e o teste final foi feito com

transações de um espaço de tempo de 10 dias, obtendo melhor resultado com um EM de 5 clusters.

Além da limitação da ausência da condução de um processo rigorosamente experimental, os dados utilizados para treinamento dos algoritmos não eram balanceados. Tão pouco os autores deixaram claro como foram feitas as rotulações das transações suspeitas, principalmente na etapa final do estudo. No experimento apresentado neste trabalho, as rotulações foram baseadas em uma investigação real e concluída, permitindo a assertividade destas.

Gao (2009) utilizou um algoritmo baseado no k-means para agrupar os dados e por meio de uma métrica encontrar as transações suspeitas. Entretanto, os dados utilizados, para etapa de testes, são gerados artificialmente e nenhuma métrica é utilizada para avaliar os resultados obtidos. O estudo apenas cita quantas transações suspeitas foram identificadas.

Umadevi et al. (2012) apresentam um estudo que usa o k-means para gerar os clusters e um minerador de padrões frequentes para encontrar as transações suspeitas. Porém, o estudo não utilizou dados reais, nem qualquer métrica para avaliar os resultados, apresentando como resultado apenas uma ferramenta de visualização.

Em outros trabalhos, o K-means é utilizado em um dos passos da geração do modelo, como é o caso em (Le Khac, 2010).

Na Tabela 2, é possível observar os classificadores e as acurácias inferidas no trabalho de Chen et al. (2014), o único que apresentou a matriz de confusão, sendo eleito para as comparações com este experimento.

As acurácias foram obtidas no treinamento em que Chen et al. (2014) utilizaram todas as transações suspeitas da base, intercaladas com subamostras de transações normais, extraídas aleatoriamente. Na tabela 3, Chen et al. (2014) aumentaram o número de amostras normais, utilizando todas as transações provenientes de um período fixo de 10 dias. Como os dados não foram balanceados, isto pode ter causado o aumento

proeminente do valor da acurácia. Além disto, Chen et al. (2014) não deixaram claro que agrupamento foi utilizado, podendo ter sido considerado o agrupamento diário ou um novo agrupamento de 10 dias.

Tabela 2. Listas dos principais testes por tipo de agrupamento, classificadores utilizados e acurácias obtidas (Chen et al., 2014).

Frequência	K-Means	EM
Dia	58,4%	81,95%
Semana	56,37%	79,75%
Mês	54,72%	79,89%

Tabela 3. Lista de acurácias no trabalho de Chen et al. (2014), por números de clusters.

Número de Clusters	EM
3	95,77%
4	96,63%
5	97,95%

Importante notar que os trabalhos utilizaram bases de dados diferentes para os seus estudos, impossibilitando uma comparação direta dos resultados. Ademais, outro fator significativo é o fato de que nenhum deles ter sido aplicado no contexto de um LABLD.

Vale ressaltar novamente que o contexto do experimento aqui apresentado vai além dos trabalhos citados, tratando-se de dados de uma investigação criminal real. Em conclusão e além do contexto, não foram encontrados trabalhos que realizassem uma análise comparativa de algoritmos aplicados ao contexto de lavagem de dinheiro, considerando uma abordagem experimental, com a validação estatística da significância dos dados, como é proposto neste trabalho.

De fato, uma base de conhecimento robusta só poderá ser gerada com as replicações de verdadeiros experimentos controlados que validem estatisticamente seus trabalhos, as quais poderão servir de insumo para verdadeiras metanálises dos dados. Este experimento tenta contribuir com a construção desta base.

5.2 Definição do Objetivo

O objetivo deste estudo foi avaliar os algoritmos EM e K-means, identificando o melhor algoritmo em termos eficácia, com foco na detecção de transações suspeitas, no âmbito de uma investigação criminal do LAB-LD de Sergipe.

O experimento terá como alvo as transações investigadas em um caso de Lavagem de Dinheiro. O objetivo foi formalizado utilizando o modelo GQM (Goal Question Metric) proposto por [1]: **Analisar,** por meio de experimento controlado, os algoritmos EM e Kmeans aplicados ao contexto de uma investigação criminal, com foco na detecção de transações suspeitas, **com a finalidade de** avaliar o melhor algoritmo em termos de eficácia (contra resultados publicados na literatura), **com respeito** à acurácia, *Log Score e RSME*, do ponto de vista de investigadores, pesquisadores e profissionais de *Data Analytics*, no **contexto** dos dados sobre transações suspeitas de uma investigação do LAB-LD.

5.3 Planejamento

5.3.1 Seleção de Contexto

O experimento foi "in vivo" e considerou os dados de transações de uma investigação do LAB-LD de Sergipe. A seleção de dados levou em consideração atributos impessoais.

5.3.2 Formulação de Hipóteses

Para guiar o estudo, foram elaboradas as seguintes questões de pesquisa, cujas respostas visam cumprir o objetivo do trabalho:

Q1: No contexto das análises investigativas conduzidas pelos LAB-LDs de Sergipe, o algoritmo EM possui maior eficácia que o K-Means, na detecção de transações financeiras suspeitas?

Q2: As eficácias alcançadas pelos algoritmos EM e K-Means, encontradas na literatura (Chen. et. al, 2014), mantêm-se para o cenário dos LAB-LDs de Sergipe?

Para avaliar a Q1, além da Acurácia, planejada na introdução desta dissertação, foram utilizadas mais duas métricas: *Log Score e RMSE*. Sendo assim, com os objetivos e métricas definidas, será considerada a hipótese secundária a seguir (para cada métrica):

H0: Os algoritmos possuem a mesma média da métrica.

$\mu 1(\text{métrica}) = \mu 2(\text{métrica}) \ldots = \mu n(\text{métrica})$;

H1: Os algoritmos possuem médias da métrica distintas.

$\mu 1(\text{métrica}) \neq \mu 2(\text{métrica}) \ldots \neq \mu n(\text{métrica})$;

5.3.3 Seleção de Participantes

Foram consideradas todas as transações de uma investigação criminal de um LAB-LD de Sergipe. Esta investigação selecionada já foi finalizada. A base de dados analisada foi coletada por meio do SIMBA (Sistema de Investigação de Movimentações Bancárias), que foi criado para facilitar e agilizar o recebimento, o processamento e a análise de informações bancárias. Este sistema foi desenvolvido pela Assessoria de Pesquisa e Análise (ASSPA) do Ministério Público Federal, sendo utilizado por diversas instituições, tais como os Ministérios Públicos dos Estados e da União, a Receita Federal do Brasil, a Polícia Federal e as Polícias Civis dos Estados e do Distrito Federal (MPMG, 2017).

A instituição cedeu a base de dados, desprovida de dados que pudessem identificar os investigados para o experimento em questão. A base de dados continha 667 transações investigadas, com 20 atributos. Os atributos selecionados são descritos na seção 5.3.4.

5.3.4 Variáveis independentes

Para a tarefa de classificação, foram considerados 14 atributos, que podem ser observados na Tabela 4. Desses 14 atributos, 8 foram criados na etapa de seleção e engenharia de atributos (vide seção de Preparação), por meio dos agrupamentos de dados por periodicidade (Dia, Mês e Ano).

Tabela 4. Atributos considerados para a análise

Atributo	Descrição
NUMERO BANCO	Código do banco da transação
TIPO	Tipo de transação
DESCRICAO LANCAMENTO	Descrição da transação
VALOR TRANSACAO	Valor da transação
NATUREZA LANCAMENTO	Natureza do lançamento da transação
OBSERVAÇÃO	Observação relacionada à transação
MONTANTE POR DIA	Valor total corrente movimentado por dia na conta
FREQUENCIA POR DIA	Frequência de movimentação diária corrente
MONTANTE POR SEMANA	Valor total corrente movimentado por semana na conta
FREQUENCIA POR SEMANA	Frequência de movimentação semanal corrente
MONTANTE POR MÊS	Valor total corrente movimentado por mês na conta
FREQUENCIA POR MÊS	Frequência de movimentação mensal corrente
MONTANTE POR ANO	Valor total corrente movimentado por ano na conta
FREQUENCIA POR ANO	Frequência de movimentação anual corrente

Por fim, nossos principais tratamentos são os algoritmos EM e K-means, com os parâmetros apresentados na Tabela 5.

Tabela 5. Parâmetros utilizados por algoritmo.

Algoritmo	EM	K-Means
CLUSTER COUNT	10	2
CLUSTER SEED	0	0
MINIMUM SUPPORT	1	1
MODELLING CARDINALITY	10	10
STOPPING TOLERANCE	10	10
MAXIMUM STATES	100	100

Os parâmetros dos algoritmos são elucidados abaixo:

CLUSTER COUNT: Especifica o número aproximado de clusters a serem criados pelo algoritmo.

CLUSTER SEED: Especifica o número de propagação usado apenas para gerar clusters aleatoriamente, no estágio inicial de criação de modelo.

MINIMUM SUPPORT: Especifica o número mínimo de casos necessários para criar um cluster. Se o número de casos do cluster for menor que esse número, o cluster será tratado como vazio e descartado.

MODELLING CARDINALITY: Especifica o número de modelos de exemplo, construídos durante o processo de clustering.

STOPPING TOLERANCE: Especifica o valor usado para determinar quando a convergência é alcançada e o algoritmo terminou de criar o modelo. A convergência é alcançada quando a alteração geral nas probabilidades do cluster é menor do que a taxa do parâmetro STOPPING TOLERANCE, dividida pelo tamanho do modelo.

MAXIMUM STATES: Especifica o número máximo de estados de atributo para os quais o algoritmo dará suporte. Se um atributo tiver mais estados que o valor máximo, o algoritmo usará os estados mais populares e ignorará os demais estados.

Os parâmetros que influenciavam significativamente a acurácia foram ajustados para melhorar o desempenho dos algoritmos. Entretanto, os parâmetros e características utilizadas não são exatamente iguais aos de Chen et. al (2014), devido aos contextos e bases diferentes. Por exemplo, durante a etapa de *tunning* do modelo, foi observado que o EM com 10 clusters apresentava um desempenho superior, enquanto Chen et. al (2014) usaram, no máximo, 5 clusters, como foi dito na seção 5.1.

5.3.5 Variáveis dependentes

Acurácia, Log Score e RMSE.

5.3.6 Projeto do Experimento

O projeto do experimento refere-se às seguintes etapas: configuração do ambiente de desenvolvimento, ou seja, o download e a instalação de todas as ferramentas mencionados na seção 5.3.7. Posteriormente, a implementação e execução do pré-processamento de dados, seleção de atributos, treinamento e testes dos modelos. Em conclusão, a execução dos testes estatísticos para a avaliação das hipóteses definidas.

Com relação ao treinamento, foi utilizada a abordagem 10 Fold Cross-validation (Hastier, 2009), em que os dados são divididos em 10 partes, mantendo suas proporções. Assim, são realizados10 testes, nos quais uma parte dos dados é separada para ser testada posteriormente e as demais são usadas para serem treinadas.

5.3.7 Instrumentação

Para o pré-processamento de dados, foi utilizada a ferramenta *Scikit-learn* (Pedregosa et al., 2011), a qual possui diversos algoritmos de aprendizado de máquina que podem ser utilizados para extrair informações relevantes de uma base de dados, com auxílio das bibliotecas de tratamento de dados *numpy* e *pandas*. A implementação do algoritmo *Random Forest* da *Scikit-learn* foi utilizada na etapa de seleção de atributos, como descrito na seção 5.4.1.2. O algoritmo SMOTE, o qual foi utilizado para balanceamento dos dados, é proveniente da biblioteca *imbalanced-learn*.

Para a carga dos dados, houve o apoio do *SQL Server*, utilizado para criação de um ETL (*Extract, Transform and Load*) com a funções de extrair, limpar e carregar os dados em um *Data Warehouse* específico, o qual é a base para geração dos modelos de conhecimento, levando em consideração as variáveis detalhadas na Tabela 4.

O processo de execução dos algoritmos e criação dos modelos foi realizado no *SQL Server Analysis Services*, utilizando os algoritmos do ambiente de desenvolvimento estendido do Visual Studio, SSDT (SQL Server Data Tools).

5.4 Operação do Experimento

5.4.1 Preparação

A primeira etapa do projeto foi a atribuição de rótulos às transações, para que para que a eficácia dos algoritmos pudesse ser mensurada. Com um caso real finalizado, foi possível supervisionar a base, classificando, previamente e com exatidão, quais eram as transações fraudulentas.

Em seguida, foi realizado o pré-processamento, o qual consistiu na engenharia e seleção de atributos, remoção de registros com dados nulos e balanceamento das classes, pois uma das métricas utilizadas neste trabalho foi a acurácia, a qual exige o balanceamento dos dados das classes.

5.4.1.1 Seleção de Atributos

A seleção e engenharia dos atributos foi realizada após análise dos trabalhos relacionados, principalmente o de Cheng et. al. (2014), que agrupou os dados das transações por períodos de tempo: diário, semanal e mensal. Além disso, citou como atributos mais importantes: montante total de debito, montante total de crédito, frequência de debito e frequência de crédito. Sendo assim, foram criados e adicionados 8 novos atributos para computar montante e frequência dos agrupamentos de dados por período de tempo.

Os atributos disponíveis na base de dados que continham informações constantes ou incompletas foram removidos. No total, 14 atributos foram selecionados (vide Tabela 4), além de um atributo classe que define o tipo da transação (fraudulenta ou não). Esta seleção teve como base os atributos que obtiveram maior peso, calculados pelo processo de decisão do algoritmo *Random Forest*, o qual foi utilizado exclusivamente para este fim.

Depois da seleção algorítmica, um especialista do LABLD ratificou os atributos, com base na possibilidade de generalização do modelo e aplicação em outros casos. O atributo que representa o rótulo das classes, transações suspeitas, só é utilizado para avaliar as predições.

5.4.1.2 Balanceamento

Para esta etapa, foi utilizado o algoritmo de balanceamento *Sythetic Minority Oversampling Technique (SMOTE)* (CHAWLA, 2002), o qual gera novas amostras baseadas na interpolação de instâncias das classes minoritárias. Desta forma, baseado no *k nearest neighbors* (kNN), o algoritmo aleatoriamente seleciona amostras das classes minoritárias e gera as novas amostras.

5.4.1.3 Carga

Consistiu na execução do ETL implementado para carga do *Data Warehouse*.

5.4.2 Execução

Na execução do experimento, com a definição dos algoritmos e atributos, uma base de treinamento foi submetida à ferramenta de mineração de dados SQL Server Data Tools (SSDT) (SQL Server Data Tools), na qual foram gerados modelos de conhecimento com o objetivo de realizar testes dos algoritmos e comparar a eficácia.

Em suma, foi realizado o processo classificatório das transações, planejado na seção 5.3.6, utilizando o dicionário discutido na subseção 5.3.4.

5.4.3 Validação dos Dados

Como auxílio para análise, interpretação e validação, foram utilizados quatro tipos de testes estatísticos, *KolmogorovSmirnov (KS) Lilliefors*, *Shapiro-Wilk*, *Teste-T Pareado* e o não paramétrico *Wilcoxon Pareado*.

Os testes *KS Lilliefors e Shapiro-Wilk* foram utilizados para os testes de Normalidade. A versão *Lilliefors* do KS foi utilizada por apresentar desempenho superior ao KS original, para pequenas amostras.

O Teste-T Pareado foi utilizado para comparar os grupos de valores diários, semanais e mensais, que apresentaram uma distribuição de valores normal, segundo os testes KS e Shapiro-Wilk. Nos casos em que foi possível rejeitar a hipótese de que o grupo de valores apresentava uma distribuição normal, em ambos os testes de normalidade, foi utilizado o teste Wilcoxon, para fins de comparação. O teste de Wilcoxon compara as médias das amostras pareadas, verificando a magnitude da diferença.

Todos os testes estatísticos foram feitos utilizando a Ferramenta SPSS – IBM (SPSS, 2017).

5.5 Resultados

Para auxiliar nos cálculos e verificar se existiam diferenças significativas na eficácia dos algoritmos, foi utilizado a fermenta para análise de dados *Statistical Package for Social Science - SPSS* (SPSS, 2017), aplicando técnicas estatísticas básicas e avançadas. O SPSS é um software estatístico internacionalmente utilizado há muitas décadas, desde suas versões para computadores de grande porte (MUNDSTOCK, 2006).

5.5.1 Análise e Interpretação de Dados

Após a execução dos algoritmos, utilizando a abordagem 10-*Cross-validation*, foram obtidos os resultados das classificações. Na Tabela 6 e na Figura 1, são apresentadas as médias das métricas obtidas por cada algoritmo.

Tabela 6. Comparativo das métricas dos algoritmos.

Algoritmo	Agregação	Acurácia	Logscore	RMSE
EM	Diário	98,25%	-0,067	0,0608
EM	Semanal	97,16%	-0,0661	0,0726
EM	Mensal	94,62%	-0,1125	0,1177
K-Means	Mensal	84,24%	-0.376	0,17
K-Means	Semanal	65,57%	-0,5632	0,3401
K-Means	Diário	51,76%	-0,6875	0,4776

Estes resultados foram utilizados para responder à questão de pesquisa Q1. Como é perceptível, os algoritmos obtiveram médias de acurácias distintas e o algoritmo EM obteve as maiores médias. Todavia, não é possível fazer essas afirmações sem evidências estatísticas suficientemente conclusivas.

Foi definido um nível de significância de 0,05 em todo o experimento. Para análise da normalidade da distribuição dos dados, foram aplicados os testes *Kolmogorov-Smirnov* (KS) *Lilliefors* e *Shapiro-Wilk*. Os *p-values* obtidos são apresentados nas tabelas 7 e 8, nas quais, observa-se que, em ambos os testes, para todas as métricas e agrupamentos, as

distribuições do EM sempre apresentam valores acima do nível de significância adotado. A partir dessas observações, conclui-se que as distribuições são normais, nestes casos supracitados.

No caso do K-means, para o agrupamento semanal, os níveis de significância para todas as métricas foram menores, no Teste de *Shapiro-Wilk* (vide Tabela 8). A reprovação em um dos testes, ou seja, p-value abaixo do nível de significância adotado, como neste agrupamento, foi suficiente para a consideração dos dados como não normais.

Para o agrupamento diário do K-means, os níveis de significância para todas as métricas também foram menores que 0.05 em ambos os testes, exceto para o RMSE, no Shapiro-Wilk. No entanto, também para este agrupamento, todas as distribuições foram consideradas não normais. Em conclusão, para o agrupamento mensal do K-means, apenas a métrica RMSE apresentou um nível de significância menor que 0.005, para ambos os testes. Desta forma, apenas esta métrica teve sua distribuição considerada não normal, para o agrupamento mensal.

O limite inferior da significância real, apresentado na tabela 7, indica que os valores reais do *p-value* estavam acima do alcance do teste *KS Lilliefors*.

Após as avaliações de normalidade, foi aplicado o teste de *Wilcoxon Pareado*, para os casos evidenciados como não normais, e o teste T Pareado, para as demais amostras que apresentaram distribuições normais. Os resultados apresentados nas Tabelas 9 e 10, com *p-values* fortemente menores que o nível de significância adotado, confirman a evidência de diferença entre as medias, ou seja, a hipótese (H0) foi rejeitada para as 3 métricas, dentro do contexto do experimento realizado.

Tabela 7. Resultado do Teste de KS Lilliefors, para análise da normalidade dos dados. * Limite inferior da significância real.

Algoritmo	Agregação	Acurácia	LogScore	RMSE
EM	Diário	0,200*	0,200*	0,200*
EM	Semanal	0,160	0,200*	0,200*
EM	Mensal	0,200*	0,200*	0,200*
K-Means	Mensal	0,200*	0,200*	0,008
K-Means	Semanal	0,118	0,139	0,116

K-Means	Diário	0,000	0,000	0,021

Tabela 8. Resultado do Teste de Shapiro-Wilk, para análise da normalidade dos dados.

Algoritmo	Agregação	Acurácia	LogScore	RMSE
EM	Diário	0,318	0,587	0,534
EM	Semanal	0,446	0,757	0,835
EM	Mensal	0,483	0,063	0,980
K-Means	Mensal	0,874	0,533	0,010
K-Means	Semanal	0,037	0,043	0,020
K-Means	Diário	0,000	0,000	0,074

Tabela 9. *p* -values do Teste-T Pareado

	Acurácia	Log Score
Mensal	0,000	0,000

Tabela 10. *p* -values do Teste de Wilcoxon

	Acurácia	Log Score	RMSE
Diário	0,005	0,005	0,005
Semanal	0,005	0,005	0,005
Mensal	X	X	0,007

Estes resultados confirmam evidências anteriores encontradas no trabalho de Chen et al. (2014). O EM se diferencia do K-means, apresentando resultados superiores em todas as métricas mensuradas. Podemos observar, na tabela 10, que os resultados obtidos por este trabalho foram superiores em todos os agrupamentos, exceto para o K-Means diário. Isto sugere que o número de clusters utilizado neste trabalho para o EM, 10, seja mais eficaz para este contexto. Outro fator que pode explicar a melhora no desempenho dos algoritmos é o balanceamento das classes para o treinamento, não efetuado no trabalho relacionado. Desta forma, **Q2** foi respondida: as eficácias, no contexto dos

LABLDs, são melhoradas para os algoritmos EM e K-Means, excetuando apenas o K-means diário, mas este também não alcançou a melhor eficácia para este agrupamento.

Tabela 11. Comparação das acurácias obtidas entre este trabalho e a literatura.

Agrupamento	EM (Chen et al., 2014)	EM	K-Means (Chen et al., 2014)	K-Means
Diário	81,95%	98,25%	58,41%	51,76%
Semanal	79,75%	97,16%	56,26%	65,57%
Mensal	79,89%	94,62%	54,72%	84,24%

Tais resultados são reforçados quando observamos os gráficos de comparação da eficácia nas figuras 2, 3 e 4. Esses gráficos mostram o desempenho do modelo para todos os estados do atributo a ser previsto. Por exemplo, ele informa quão bem o modelo prevê tanto as transações suspeitas quanto as não suspeitas (acurácia).

O eixo x do gráfico representa a porcentagem do conjunto de dados de teste usada para comparar as previsões. O eixo y representa a porcentagem de previsões corretas. Portanto, a linha ideal é a diagonal, que mostra que em 50% dos dados o modelo prevê corretamente 50% dos casos, o valor máximo que se pode esperar.

No gráfico da figura 1 (agrupamento diário), podemos observar que o EM, representado pela linha verde, acompanha o modelo ideal, representado pela linha azul, por quase todo o conjunto de teste, afastando-se do valor ideal no fim do teste. O K-Means, representado pela linha vermelha, apresenta um desempenho muito inferior a partir dos 10%, que só melhora aproximadamente aos 57% do teste. Mas mesmo assim, o desempenho do EM se mostra muito superior no final. No gráfico da figura 2 (agrupamento semanal), podemos observar um comportamento parecido, no entanto, o EM se afasta do modelo ideal um pouco mais cedo e o K-Means também começa sua ascenção mais cedo. Já no último gráfico (agrupamento mensal), figura 3, podemos ver que o desempenho do K-Means é muito superior com relação as suas versões anteriores, com o EM mantendo quase o mesmo padrão que já havia apresentado.

É importante ressaltar que ao contrário do resto do trabalho, os dados representados nesses gráficos não passaram por um *Cross validation* com k = 10, por isso a acurácia final dos modelos pode não estar representada com exatamente os mesmos valores. Tem-se o modelo o treinado (70% da base) sendo avaliado com uma base de testes (30% da base).

Figura 1: Gráfico de comparação de precisão entre o K-Means e o EM para o agrupamento diário.

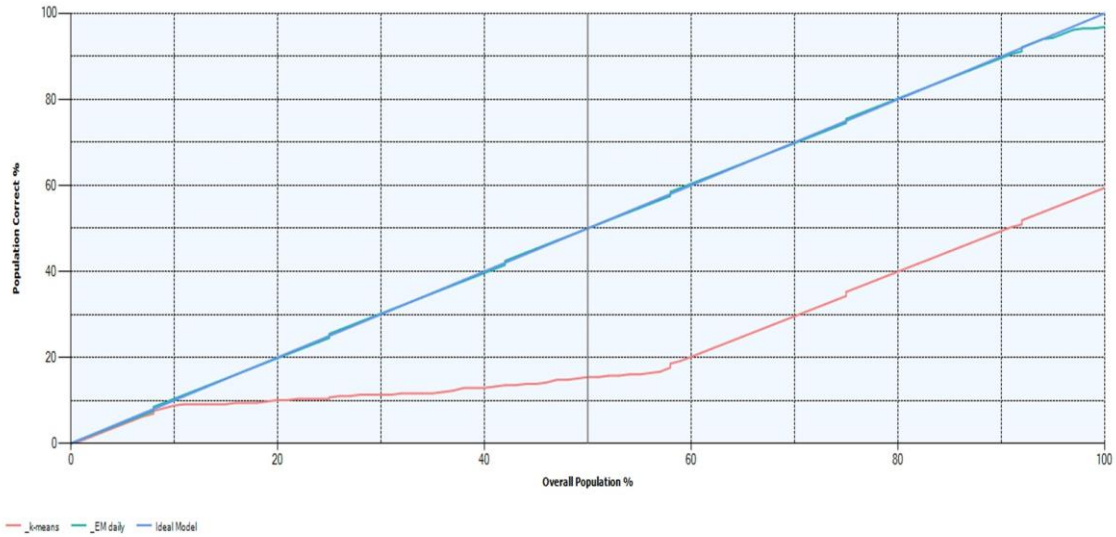


Figura 2: Gráfico de comparação de precisão entre o K-Means e o EM para o agrupamento semanal.

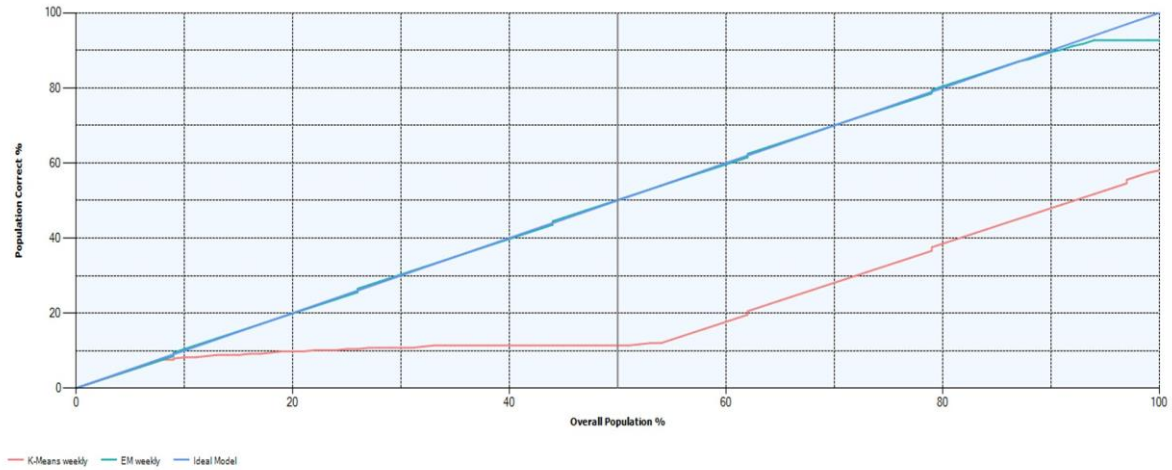
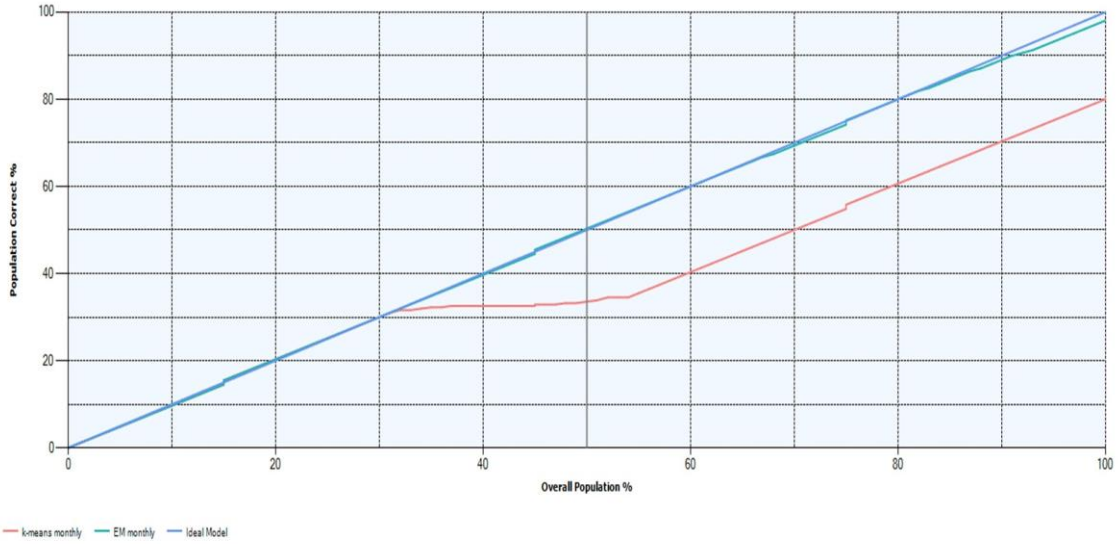


Figura 3: Gráfico de comparação de precisão entre o K-Means e o EM para o agrupamento mensal.



5.5.2 Ameaças à Validade

Embora os resultados do experimento tenham se mostrado satisfatórios, o mesmo apresenta ameaças à sua validade que devem ser comentadas.

Ameaças à validade interna: Instâncias importantes para a indução do modelo podem ter sido removidas durante o processo de limpeza dos dados, por possuírem informações faltantes. Esta ameaça foi mitigada, pois muitos valores faltantes eram de atributos com baixa correlação com a classe alvo (transação fraudulenta ou não), os quais foram eliminados e o registro aproveitado na limpeza.

Ameaças à validade de construção e à validade externa: O caso utilizado para o treinamento pode ser muito específico. Esta ameaça foi mitigada, em parte, por meio da seleção de atributos com alta correlação com a classe alvo, os quais pudessem ser generalizados para qualquer caso. De qualquer forma, mais testes devem e serão feitos para casos inéditos e em andamento, com a avaliação extrínseca da predição de transações suspeitas.

CAPÍTULO 6

6 CONCLUSÕES

As recentes implementações, pouco mais de uma década, dos Laboratórios de Tecnologia de Combate à Lavagem de Dinheiro - LABLDs, culminando com o surgimento da Operação Lava Jato em todo o país, tem posto a eficiência desta essencial iniciativa à prova. Sem a capacidade de avaliação e questionamento desta eficiência, nas operações como a Lava Jato, esta dissertação propõe apenas o início de pesquisas que incrementem os trabalhos, extrapolando o empirismo e conectando modelos científicos a estas unidades de investigação.

A utilização de modelos científicos estimula e potencializa a atuação dos LABLDs, proporcionando assim um embasamento para as análises e tomada de decisão de analistas, agentes de polícia, delegados de polícia, promotores de justiça e até mesmo juízes de direito.

Assim, esta dissertação não deve servir apenas como requisito para a aquisição de titulação acadêmica, também deve contribuir para o início de uma série de pesquisas neste campo do conhecimento, direcionando a atuação investigativa no Brasil.

6.1 Contribuições

A principal contribuição deste trabalho foi a avaliação experimental dos principais algoritmos de aprendizado não supervisionado utilizados no contexto da LD.

Além disso, outras contribuições foram:

- Um *Survey* aplicado aos principais órgãos de combate ao crime organizado, publicado no XIII Simpósio Brasileiro de Sistemas de Informação - Lavras/MG - SBSI 2017, Qualis B2.

- Mapeamento Sistemático utilizado para identificar e sistematizar as principais abordagens, técnicas e algoritmos utilizados na Ciência da Computação (CS) para combater a Lavagem de Dinheiro. Ressalte-se que os resultados serão publicados no *16th International Conference on Information Technology : New Generations - ITNG 2019*, Qualis B1, com apresentação em Las Vegas;

- A criação de um ETL e um projeto de banco de dados para automatizar o tratamento dos dados brutos disponibilizados pelo SIMBA.

Como consequência destas contribuições, conseguimos responder às questões de pesquisa elaboradas no início do trabalho:

- Q1: No contexto das análises investigativas conduzidas pelos LAB-LDs de Sergipe, o algoritmo EM possui maior eficácia que o K-Means, na detecção de transações financeiras suspeitas? Sim. O experimento demonstrou que os algoritmos obtiveram médias distintas e o algoritmo EM obteve as melhores médias

- Q2: As eficácias alcançadas pelos algoritmos EM e K-Means, encontradas na literatura, mantêm-se para o cenário dos LAB-LDs de Sergipe? Em termos de superioridade do EM, sim, os resultados encontrados confirmam as evidências detectadas no artigo base, no entanto, as eficácias foram melhores, no LABLD, para os dois algoritmos, superando os achados da literatura.

6.2 Limitações

O presente estudo apresentou uma limitação quanto ao acesso aos dados para análise, pois esta é uma área de natureza investigativa e sigilosa.

6.3 Perspectivas

O volume de dados analisados em investigações dentro das unidades dos LABLDs pode proporcionar a criação de modelos preditivos robustos e consolidados, preenchendo lacunas científicas dos trabalhos publicados na literatura. Outros possíveis desdobramentos são:

- Elaboração de metodologias padronizadas para análise de dados em casos concretos e investigações reais ocorridas nos LABLDs;
- Confecção de sistemas ou módulos auxiliares que se utilizam de modelos preditivos tal como o criado nesta dissertação, os quais possam ser utilizados automaticamente pelos por investigadores. No início de uma investigação, será possível aumentar a eficiência das buscas, garimpando transações que possuem maior probabilidade de suspeição.

6.4 Considerações Finais

Este trabalho converge positivamente para o combate à LD. A apropriação de métodos formais de análises, que doravante se disponibilizarão a auxiliar os LABLDs em todo o país, contribuirá para formação de unidades investigativas de excelência.

CAPÍTULO 7

7 REFERÊNCIAS

ALEXANDRE, Claudio; BALSA, João. Integrating client profiling in an anti-money laundering multi-agent based system. In: **New Advances in Information Systems and Technologies**. Springer, Cham, 2016. p. 931-941.

APPOLINÁRIO, F. **Dicionário de metodologia científica: um guia para a produção do conhecimento científico**. São Paulo: Atlas, 2007.

ABRAMOVICI, Michael et al. Competing fusion for bayesian applications. In: **Proceedings of IPMU**. 2008. p. 379.PIMENTEL, M.; FUKS, H. Sistemas colaborativos. Rio de Janeiro: Elsevier, 2012.

BASILI, V., Trendowicz, A., Kowalczyk, M., Heidrich, J., Seaman, C., Münch, J., Rombach, D. **Aligning Organizations Through Measurement:The GQM+Strategies Approach**, Springer, 2014.

BASILI, Victor R.; WEISS, David M. A methodology for collecting valid software engineering data. **IEEE Transactions on software engineering**, n. 6, p. 728-738, 1984.

BRASIL. Lei nº 9.883, de 07 de Dezembro de 1999. **Institui o Sistema Brasileiro de Inteligência, cria a Agência Brasileira de Inteligência – ABIN**, e dá outras providências.

BRASIL. Ministério da Justiça. Secretaria Nacional de Segurança Pública. **Doutrina Nacional de Inteligência de Segurança Pública**. Brasília, 2014.

BRAMER, M. 2007. **Principles of data mining.** Springer London, New York, NY.

COSTA, Arthur Trindade; LIMA, Renato Sergio de. **Crime, polícia e justiça no Brasil**. Organização Renato Sérgio de Lima, José Luiz Ratton e Rodrigo Ghiringhelli de Azevedo. – São Paulo: Contexto, 2014.

CHEN, Hsinchun, et al. **Crime data mining: a general framework and some examples**. Computer 37.4: 50-56, 2004.

CHEN, Zhiyuan, et al. **Exploration of the effectiveness of expectation maximization algorithm for suspicious transaction detection in anti-money laundering.** Open Systems (ICOS), 2014 IEEE Conference on. IEEE, 2014.

CHAWLA, Nitesh V. et al. SMOTE: synthetic minority over-sampling technique. **Journal of artificial intelligence research,** v. 16, p. 321-357, 2002.

DUHART, Bronson Amauri Montoya; HERNÁNDEZ-GRESS, Neil. Review of the Principal Indicators and Data Science Techniques Used for the Detection of Financial Fraud and Money Laundering. In: **Computational Science and Computational Intelligence (CSCI), 2016 International Conference on**. IEEE, 2016. p. 1397-1398.

DEMPSTER, Arthur P.; LAIRD, Nan M.; RUBIN, Donald B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the royal statistical society. Series B (methodological)**, p. 1-38, 1977.

DRESCH, A.; LACERDA, D. P.; ANTUNES JÚNIOR, J. A. V. **Design science research: método de pesquisa para avanço da ciência e tecnologia**. Porto Alegre: Bookman Editora, 2015.

ESLAMNEZHAD, Mohsen; VARJANI, Ali Yazdian. Intrusion detection based on MinMax K-means clustering. In: **Telecommunications (IST), 2014 7th International Symposium on**. IEEE, 2014. p. 804-808.

FIGUEIRA, Marcelle Gomes. **A construção de um sistema nacional de informações em segurança pública: os desafios de implementação de uma agenda**, 2015.

GAO, Shijia; XU, Dongming. **Conceptual modeling and development of an intelligent agent-assisted decision support system for anti-money laundering**. Expert Systems with Applications, v. 36, n. 2, p. 1493-1504, 2009.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining: concepts and techniques.** Elsevier, 2011.

HYNDMAN, Rob J.; KOEHLER, Anne B. **Another look at measures of forecast accuracy. International journal of forecasting,** v. 22, n. 4, p. 679-688, 2006.

JUN, Tang. A cross datasets referring outlier detection model applied to suspicious financial transaction discrimination. In: **Intelligence and Security Informatics**. Springer, Berlin, Heidelberg, 2006. p. 58-65.

JURISTO, Natalia; MORENO, Ana M. S**oftware engineering experimentation**. Springer Science & Business Media, 2001.

KIMBALL, Ralph; ROSS, Margy. The data warehouse toolkit: **the complete guide to dimensional modeling**. John Wiley & Sons, 2011.

KITCHENHAM, B. **Procedures for performing systematic reviews**. Keele, UK, Keele University, v. 33, n. TR/SE-0401, p. 28, 2004. ISSN 13537776.

LV, Lin-Tao; JI, Na; ZHANG, Jiu-Long. A RBF neural network model for anti-money laundering. In: **Wavelet Analysis and Pattern Recognition, 2008. ICWAPR'08. International Conference on**. IEEE, 2008. p. 209-215.

LE KHAC, Nhien An; KECHADI, M.-Tahar. Application of data mining for anti-money laundering detection: A case study. In: **Data Mining Workshops (ICDMW), 2010 IEEE International Conference on**. IEEE, 2010. p. 577-584.

MCAFEE, Andrew et al. Big data: the management revolution. **Harvard business review**, v. 90, n. 10, p. 61-67, 2012.

MUNDSTOCK, Elsa et al. Introdução à Análise Estatística utilizando o SPSS 13.0. **Cadernos de Matemática e Estatística Série B. Universidade Federal do Rio Grande do Sul,** Porto Alegre, RS, 2006.

PAULA, Alexandre Vagtinski de. **Determinação de parâmetros que caracterizam o fenômeno da biestabilidade em escoamentos turbulentos.** 2013.

PIMENTEL, M.; FUKS, H. **Sistemas colaborativos**. Rio de Janeiro: Elsevier, 2012.

SANTOS, R. ; MANGUEIRA, F. ; OLIVEIRA, M. ; COLAÇO JÚNIOR, Methanias. **A Survey on the use of Data Mining and Data Analytics techniques by Brazilian criminal investigation agencies**. In: Brazilian Symposium on Information Systems, 2017, Lavras. SBSI, 2017.

SANTOS, Rachel Boba. **Crime analysis with crime mapping. Sage Publications**, 2012.

SIMBA. **https://asspaweb.pgr.mpf.gov.br/site/ simba/**. accessed: 2017-12-01.

SCHOTT, P.A.: **Reference Guide to Anti-Money Laundering and Combating the Financing of Terrorism: Second Edition and Supplement on Special Recommendation IX.** The World Bank and The International Monetary Fund, Washington DC, sec- ondedn. (2006).

SEVERINO, A. J.. **Metodologia do trabalho científico**. Cortez editora, 2017.

TREVOR, Hastie; ROBERT, Tibshirani; JH, Friedman. **The elements of statistical learning: data mining, inference, and prediction**. 2009.

WANG, Xiaoyan; BAI, Yanping. A modified MinMax-means algorithm based on PSO. **Computational intelligence and neuroscience**, v. 2016, 2016.

WITTEN, Ian H. et al. Data Mining: **Practical machine learning tools and techniques**. Morgan Kaufmann, 2016.

WOHLIN, C et al.. **Experimentation in Software Engineering**, Kluwer Academic Publishers, 2012.

UMADEVI, P.; DIVYA, E. **Money laundering detection using TFA system**. 2012.

ZHU, Tianqing. An outlier detection model based on cross datasets comparison for financial surveillance. In: **Services Computing, 2006. APSCC'06. IEEE Asia-Pacific Conference on**. IEEE, 2006. p. 601-604.

ZHANG, Zhongfei Mark; SALERNO, John J.; YU, Philip S. Applying data mining in investigating money laundering crimes. In: **Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining**. ACM, 2003. p. 747-752.